

ABSTRACT

Title of dissertation: F-TYPE LECTINS: BIOCHEMICAL, GENETIC, AND
TOPOLOGICAL CHARACTERIZATION OF A NOVEL
LECTIN FAMILY IN LOWER VERTEBRATES

Eric W. Odom-Crespo, Doctor of Philosophy, 2004

Dissertation directed by: Professor Gerardo R. Vasta
Center of Marine Biotechnology
University of Maryland Biotechnology Institute

In vertebrates, immunoglobulins have long been considered the primary means for recognition of potential pathogens since they are the product of an adaptive immune system capable of generating and selecting for the most efficient antibody when adequately primed. However, initiating a naive system requires signals from cells that employ invariant receptors rather than antibodies. These innate receptors appear to recognize repetitive polymers commonly found only in microbes. Frequently, these receptors are lectins specific for polysaccharides ubiquitous of the microbial surface. Lectins in the blood or lymph are widespread among metazoans while antibodies are a vertebrate innovation suggesting that lectins may be evolutionarily their functional precursors. Even in primitive jawed vertebrates, there is a complete adaptive immune system, but it is relatively inefficient in comparison to mammals. Therefore, lectins might have a prominent immune function in lower vertebrates comparable

to antibodies. To test this, a teleost was surveyed for humoral and hepatic lectins. A fucose-specific lectin of 32 kDa (FBP32) was initially purified from the palmetto bass and upon sequencing indicated it was unlike other reported lectins. The primary structure is characterized by a tandem polypeptide motif (FBPL) with partial homology to a long pentraxin from a frog. An inflammatory challenge of bass to test if FBP32 behaved like a mammalian pentraxin indicated that the FBP32 transcript level increases, but protein levels appear constitutive. An extensive search using both molecular cloning and gene database queries revealed that FBP32 is a member of a diverse protein family reflecting varying concatenations of the FBPL and even present in a cell surface receptor, but of sporadic phylogenetic distribution most notably being absent in mammals. Analysis of FBP32s genic structure reveals that it is flanked by phase 1 introns, which may explain the domains ability to concatenate and shuffle to form mosaic proteins. In collaboration with experts, the tertiary structure of an FBPL including its fucose-binding site was elucidated revealing a novel lectin fold, the F-type lectin fold (FTL), that is shared by unrelated proteins. Characterization of FBPLs demonstrates that the study of mammals alone may not reveal the full extent of immune system innovation.

F-TYPE LECTINS: BIOCHEMICAL, GENETIC, AND TOPOLOGICAL
CHARACTERIZATION OF A NOVEL LECTIN FAMILY IN LOWER VERTEBRATES

By

Eric W. Odom-Crespo

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2004

Advisory Committee:

Professor Gerardo R. Vasta, Chair
Dr. Frederick J. Cassels
Professor Martin F. Flajnik
Professor Russell T. Hill
Professor William R. Jeffery, Dean's representative for College of Life Sciences
Professor Allen R. Place

©Copyright by
Eric W. Odom-Crespo
2004

DEDICATION

I dedicate this dissertation to my wife, Sandra, and our two little joys, which are soon to be born. I could not have done this without you, Sata. I would not have continued this effort if it were not for the example set by my mother, Ligia Milagros Crespo-Odom, who persevered in spite of much suffering. Finally, my father, Sanders Odom, Jr., whose strength of character and tenderness towards my ailing mother taught me what it means to be a husband.

ACKNOWLEDGMENTS

I would like to thank the financial support provided by the NIGMS through their pre-doctoral fellowship (GM14903-04). Thanks to Dr. Yonathan Zohar for generously providing the *M. saxatilis* genomic DNA library. Thanks also to Dr. Michael S. Quesenberry and other lab members for their respective contributions to this project. Finally, I want to acknowledge the patience and encouragement provided by my advisor, Dr. Gerardo R. Vasta.

TABLE OF CONTENTS

Dedication.....	ii
Acknowledgments.....	iii
Table of contents.....	iv
List of tables.....	viii
List of figures.....	ix
List of abbreviations.....	xii
Common names of species	xvi
Chapter 1. Introduction.....	1
A synopsis of adaptive immunity.....	1
The immunoglobulin fold.....	3
A synopsis of innate immunity	4
Cellular innate receptors	6
Acute phase proteins	7
Lectins.....	8
Immunity in ancient vertebrates	12
Fish as models of immunity	15
Project Goal and Significance	15
Chapter 2. Identification of a unique bass lectin	17
Materials and Methods.....	19
Reagents.....	19
Animals	19
Extraction of blood and separation of serum	20
Dissection and collection of livers.....	20
Preparation of liver extracts.....	20
Protein concentration estimation.....	22
Protein purification and storage.....	22
A. Preparation of affinity chromatography gel.....	22
B. Affinity chromatography of liver acetone extracts.....	23
C. Anion exchange batch chromatography (AEX) of aqueous liver extract	24
D. Affinity chromatography of aqueous liver extracts	24
E. Affinity chromatography of serum	25
F. Size exclusion chromatography	25
G. Storage of purified bass lectin	26
Analytical procedures	26
A. Electrophoresis and isoelectric focusing.....	26
B. Glycosylation analysis	26
C. Electron spray ionization-mass spectrometry (ESI-MS)	27
D. Size exclusion chromatography	27
E. Chemical crosslinking.....	28
Carbohydrate-binding properties	29
A. Hemagglutination assay.....	29
B. Hemagglutination inhibition	30

C. Cation requirements	30
Peptide and nucleic acid sequence	31
A. Edman sequence analysis.....	31
B. Isolation of total RNA from bass liver.....	32
C. Synthesis of first strand cDNA	34
D. cDNA cloning	35
E. Completion of full length cDNA.....	36
F. DNA sequencing.....	37
G. Bioinformatics	37
Immune challenges.....	37
A. Northern analysis	37
B. Western analysis	39
Results and Discussion	41
Purification of a fucose-binding lectin	41
Characterization of the native FBP32	48
Isoelectric focusing analysis.....	49
Carbohydrate specificity.....	50
Cation requirements	56
Cloning of the hybrid bass lectin	56
Expression analysis of FBP32.....	63
Chapter 3. Emergence of a novel lectin family.....	66
Materials and Methods.....	66
Cloning FBP32 cDNA from parental species.....	66
Gene database search for homologues	67
Cloning and sequencing of <i>Xenopus laevis</i> FBPL ESTs	68
Design of FBPL degenerate primers.....	70
Detection of FBPLs by MOPAC	71
Sequencing of zebrafish FBPL ESTs	71
Sequencing of steelhead trout FBPL EST	71
Sequencing of fruit fly FBPL EST.....	73
Analysis of expression of CG9095 during fruit fly development.....	75
Multiple alignment of peptide sequences.....	75
Phylogenetic analysis	76
Results and Discussion	76
Meeting the parents	76
A unique lectin family	80
Revision of PXN1-XENLA.....	81
The march of FBPLs in tetrapods	85
Diversity of FBPLs from ray-fined fish	91
Fish can count.....	99
FBPLs of invertebrates	106
Paucity of FBPs in prokaryotes: vertical vs. horizontal transfer?.....	113
FBPL motif and domain topology	119
Phylogenetic analysis of binary FBPLs.....	128
Unknown fate of FBPs during tetrapod evolution.....	131
Displacement of FBPLs by C-type lectins?	136

Chapter 4. Genetic analysis of FBPL family	138
Materials and Methods.....	138
Screening of striped bass genomic DNA bacteriophage library	138
Southern blot analysis.....	141
PCR amplification of MsaFBP32 gene from λ Ms15.....	141
Deletional directional sequencing.....	142
Bacteriophage clone shotgun sequencing	142
Cloning of MsaFBPII cDNA from liver	145
PCR amplification of Mcfbp32.....	145
Isolation of genomic DNA.....	146
PCR amplification of FBP32 gene in hybrid bass.....	146
Screening of zebrafish genomic DNA bacteriophage library	147
PCR amplification of cDNA encoded by λ Zf13.....	148
Genome sequence database searches	148
Global alignments and comparative search for regulatory regions.....	149
Results and Discussion	149
A domain divided	149
Genomics of FBP in the zebrafish model	161
A view of the whole fish.....	165
Searching for the ON switch.....	169
Chapter 5. Structural elucidation of an FBPL.....	173
Materials and Methods.....	174
Repurification of AAA from a commercial preparation.....	174
Size exclusion chromatography.....	174
Crystallization	174
X-ray diffraction data collection	175
Structure determination	175
Results and Discussion	176
Crystallization of AAA.....	176
A novel lectin fold	176
Saccharide-binding site.....	178
Calcium-binding site.....	183
Quaternary structure of AAA	185
A structural view of the FBPL motif.....	188
Structural analogy is widespread	193
Conclusions and future directions.....	199
Appendices.....	206
A. MsaFBP32 cDNA.....	206
B. Xla-PXN1 cDNA.....	208
C. XlaII-FBPL cDNA.....	212
D. Xla-neurula cDNA	215
E. <i>In silico</i> XLEST2 cDNA	217
F. DreI-FBPL cDNA.....	219
G. DreII-FBPL partial CDS.....	221
H. DreIII-FBPL CDS	222

I. DreIV-FBPL <i>in silico</i> CDS.....	224
J. DreV-FBPL cDNA.....	226
K. Omy-FBPL4 cDNA.....	228
L. CG9095 cDNA.....	231
M. Pairwise alignment of MsaFBP32 and MchFBP32 genes.	234
Bibliography.....	244

LIST OF TABLES

Table 1. Purification table of the bass Fuc-binding lectin from serum.....	47
Table 2. Hemagglutination inhibition profile of hybrid bass FBP32 by monosaccharides, oligosaccharides, and glycoproteins *	55
Table 3. Edman sequencing of bass hepatic lectin.....	57
Table 4. Predicted intron splice sites for DreIV-FBPL gene	98
Table 5. Taxonomic distribution and properties of deduced polypeptides containing an FBPL.....	121
Table 6. Predicted intron splice sites of <i>F. rubripes</i> FBPL genes.	168

LIST OF FIGURES

Figure 1. SDS-PAGE analysis of the hybrid bass hepatic lectin.	42
Figure 2. SEC polishing of Fuc affinity-purified serum lectin.....	43
Figure 3. SDS-PAGE analysis of the purification steps for the bass serum lectin.	43
Figure 4. SDS-PAGE analysis of the serum form of the Fuc bass lectin.	44
Figure 5. Determination of MsaFBP32 mass by ESI-MS.....	45
Figure 6. Size estimation of MsaFBP32 by SEC.....	48
Figure 7. SDS-PAGE of MsaFBP32 cross-linked with BS3.	49
Figure 8. Denaturing isoelectric focusing of MsaFBP32.....	50
Figure 9. 2-D PAGE analysis of SEC-grade MsaFBP32.....	50
Figure 10. Hemagglutination inhibition curves for A. monosaccharides, B. polysaccharides and C. glycoproteins.....	54
Figure 11. PCR amplification of MsxMcFBP32 transcript by MOPAC.....	58
Figure 12. 5' and 3' RACE of MsxMcFBP32 cDNA.....	58
Figure 13. Sequencing contig scheme of the full-length MsxMcFBP32 cDNA.....	58
Figure 14. The complete cDNA and deduced protein sequence of MsxMcFBP32.	61
Figure 15. Duplicated motif present in the deduced sequence of MsxMcFBP32.....	62
Figure 16. Immunoblot detection of MsaFBP32 during purification.....	64
Figure 17. Northern analysis of FBP32 in liver after turpentine challenge.	65
Figure 18. Pairwise alignment of MsaFBP32 and MchFBP32 full-length cDNAs.....	78
Figure 19. Northern analysis of striped bass liver RNA probed with FBP32 probe.....	79
Figure 20. Northern blot detection of tissue expression of MsaFBP32.....	80
Figure 21. Similarity of the <i>Xenopus laevis</i> pentraxin-fusion protein to the tandem domains of the bass lectin.	81
Figure 22. <i>X. laevis</i> liver EST matching the 5' end of published PXN1-XENLA cDNA (L19881).....	82
Figure 23. 5' RACE of XL-PXN1.....	83
Figure 24. PCR amplification of full length Xla-PXN.	84
Figure 25. Sequencing contig of full length Xla-PXN.	84
Figure 26. Support for completed ORF of Xla-PXN.....	85
Figure 27. <i>In silico</i> assembly <i>X. laevis</i> liver EST contig unique from Xl-PXN.....	85
Figure 28. Partial sequencing path of purchased <i>X. laevis</i> liver cDNA clones.....	86
Figure 29. PCR amplification of full-length XlaII-FBPL.....	87
Figure 30. Sequencing contig of the liver XlaII-FBPL containing four FBPL domains.....	87
Figure 31. Multiple alignment of <i>X. laevis</i> neurula ESTs encoding single FBPL (XlaEST2)	88
Figure 32. <i>In silico</i> assembled contig (Xla-neurula) from <i>X. laevis</i> embryos encoding a single FBPL.....	88
Figure 33. <i>In silico</i> assembled contig of XtrII-FBPL.....	89
Figure 34. <i>In silico</i> assembled contig of XtrIII-FBPL.....	90
Figure 35. Design of degenerate DNA primers for MOPAC using CODEHOP.	92
Figure 36. PCR amplification from genomic DNA of a zebrafish FBPL homologue by MOPAC.....	93

Figure 37. Nested 3' RACE of DreI-FBPL.....	93
Figure 38. Nested 5' RLM-RACE of DreI-FBPL.....	94
Figure 39. Completion of DreI-FBPL sequence.....	94
Figure 40. 3' RACE with nested DNA primers of DreII-FBPL identified from λ Zf13.....	95
Figure 41. PCR Amplification of DreIII-FBPL with gonadal EST-specific primers.....	96
Figure 42. 5' RACE amplicons of trout FBPL from liver total RNA.....	99
Figure 43. Colony PCR of OmyFBPL4 5' RACE subclones.....	100
Figure 44. Contig of RACE amplicons for OmyFBPL4.....	100
Figure 45. PCR amplification of full-length OmyFBPL4.....	101
Figure 46. Sequencing contig of 1 kb chimera.....	102
Figure 47. Phylogeny of living fish [238].....	105
Figure 48. 5' RACE of LP08801 from larval cDNA library lysates.....	107
Figure 49. Sequencing contig of the full length <i>D. melanogaster</i> CG9095.....	107
Figure 50. Goldman-Engelman-Steitz hydrophobicity plot of the mature CG9095.....	108
Figure 51. RT-PCR amplification of CG9095 from fruit fly embryonic, larval and adult RNA.....	109
Figure 52. Alignment of the segments from CG9095 and <i>furrowed</i> spanning the FBPL and CTLD from the <i>Drosophila</i> homologues.....	111
Figure 53. Schematic representation of domain organization by ProDom for <i>S. pneumoniae</i> SP2159 (above) and <i>C. perfringens</i> CPE0329 (below) proteins.....	116
Figure 54. Domain architecture of the bass FBP32 homologues.....	122
Figure 55. Alignment of the FBPL domain.....	126
Figure 56. Analysis of genetic distance between FBPL domains.....	127
Figure 57. Analysis of evolutionary distance between binary FBPLs.....	130
Figure 58. Tentative identification of an FBPL sequence in the ascidian, <i>C. intestinalis</i>	135
Figure 59. Restriction enzyme mapping of striped bass' genomic DNA clones.....	150
Figure 60. Autoradiogram of restriction analysis of λ Ms bacteriophage clones.....	151
Figure 61. Sequencing contig path of directionally deleted subclones of pMs15s.1.....	151
Figure 62. Structure of FBP32 cDNA and gene.....	154
Figure 63. Tc1/ <i>mariner</i> -like (TLE) transposons in FBP32 introng.....	155
Figure 64. Contig path of directionally-deleted pMc3.....	156
Figure 65. Detection in palmetto hybrid of both FBP32 genes from parental species.....	156
Figure 66. Sequencing contig of full λ Ms15 insert. The lowest arrow is that of pMs15s.1.....	157
Figure 67. Map of λ Ms15 indicating the position of FBP genes and the transposon-like elements within.....	158
Figure 68. Comparison of MsaFBP cDNAs.....	159
Figure 69. Overlap of λ Ms bacteriophage clones.....	160
Figure 70. Scheme of the fruit fly CG9095 cDNA.....	161
Figure 71. Cloning of DreI-FBPL gene.....	162
Figure 72. Cloning of DreII-FBPL gene, a second homologue from zebrafish.....	164
Figure 73. Identification of two pairs of orthologous FBP genes in pufferfish.....	167
Figure 74. Alignment (rVISTA) of first exon and region upstream for MsaFBPII and MsaFBP32.....	170
Figure 75. Global alignment of <i>F. rubripes</i> scaffold_237 to <i>T. nigroviridis</i> scaffold 461_3b.....	171
Figure 76. Global alignment of <i>F. rubripes</i> scaffold_599 to <i>T. nigroviridis</i> scaffold 794_4b.....	172

Figure 77. F-type lectin fold.	178
Figure 78. Sugar-binding site.	181
Figure 79. Heptacoordinated Ca^{2+} -binding site of AAA.	184
Figure 80. Quaternary structure of AAA.	188
Figure 81. Conservation of functional positions within the FBPL family.	192
Figure 82. AAA shares structural similarity with diverse proteins.	194
Figure 83. Structure-based sequence alignment of AAA and its structural analogs.	197

LIST OF ABBREVIATIONS

2-ME	2- β -mercaptoethanol
AAA	<i>Anguilla anguilla</i> agglutinin
AEX	anion exchange
APC	antigen-presenting cell
APP	acute phase protein
APR	acute phase reaction
ATCC	American Type Culture Collection
AU	agglutination units
BLAST	basic local alignment research tool
bp	base pair
CBM	carbohydrate-binding module
CCP	complement control protein
cDNA	complementary deoxyribonucleic acid
CDR	complementarity determining region
CDS	coding sequence
CIP	calf intestine phosphatase
CMR	Comprehensive Microbial Resource
CPS	capsular polysaccharides
CRP	C-reactive protein
CTLD	C-type lectin domain
dbEST	Genbank EST database
DC	dendritic cell
DEAE	diethylaminoethyl
D-Fuc	6-deoxy-D-galactose
DNA	deoxyribonucleic acid
DS	discoidin domain
EDTA	ethylenediaminetetraacetic acid
ESI-MS	electron spray ionization-mass spectrometry
EST	expressed sequence tag

FA58C	coagulation factor 5 and 8 c domain
FBP	fucose-binding protein
FBPL	fucose-binding protein-like
FTLD	F-type lectin domain
Fuc	6-deoxy-L-galactose
Gal	D-galactose
Glc	D-glucose
GlcNAc	N-acetyl glucosamine
HGT	horizontal gene transfer
HRPO	horse radish peroxidase
IACUC	Institutional Animal Care and Use Committee
IC50	inhibitor concentration at 50% inhibition
IEF	isoelectric focusing
Ig	immunoglobulin domain
IgM	immunoglobulin μ
Igsf	immunoglobulin superfamily
IPTG	isopropyl-1-thio- β -D-galactoside
kb	kilobase
kt	electric potential
Lac	D-galactose β (1-4) D-glucose
LPS	lipopolysaccharide
Man	D-mannose
MBP	mannan-binding protein
MHC	major histocompatibility complex
mIg	membrane ig
MOPAC	multiple oligonucleotide PCR amplification of cDNA
MPD	2-methyl-2,5-pentenediol
Mya	million years ago
NBT/BCIP	nitroblue tetrazolium/ 5-bromo-4-chloro-3-indolyl phosphate
oligo(dT)	oligodeoxythymidylic acid
ORF	open reading frame

PAMP	pathogen-associated molecular patterns
PCR	polymerase chain reaction
pfu	plaque-forming units
PG	peptidoglycan
pI	isoelectric point
PMSF	phenylmethylsulfonyl fluoride
poly(A)+	polyadenylated (mRNA)
PVDF	polyvinylidene difluoride
RACE	rapid amplification of cDNA ends
RAG	rearrangement-associated genes
RBC	red blood cells
RNA	ribonucleic acid
RT-PCR	reverse transcription/polymerase chain reaction
SA	specific activity
SAA	serum amyloid A
SCOP	structural classification of proteins database
SCR	short consensus repeats
SDS-PAGE	sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SEC	size-exclusion chromatography
TA	teichoic acid
TAE	tris/acetate/edta electrophoresis buffer
Taq	<i>Thermus aquaticus</i> DNA polymerase
TBS	tris-buffered saline
TCR	T-cell receptor
TD	thymus-dependent antigens
TFBS	transcription factor-binding site motifs
TI	thymus-independent antigens
TLE	Tc1/ <i>mariner</i> -like element
TLR	Toll-like receptor
TSS	transcription start site
UTR	untranslated mRNA region

V	variable immunoglobulin domain
β -gal	β -galactosidase

COMMON NAMES OF SPECIES

African clawed frog	<i>Xenopus laevis</i>
Channel catfish	<i>Ictalurus punctatus</i>
Chinook salmon	<i>Oncorhynchus tshawytscha</i>
Common carp	<i>Cyprinus carpio</i>
Eastern oyster	<i>Crassostrea virginica</i>
European eel	<i>Anguilla anguilla</i>
European sea bass	<i>Dicentrarchus labrax</i>
Fruit fly	<i>Drosophila melanogaster</i>
Goldfish	<i>Carassius auratus</i>
Green spotted pufferfish	<i>Tetraodon nigroviridis</i>
Harlequin rasbora	<i>Rasbora heteromorpha</i>
Japanese eel	<i>Anguilla japonica</i>
Japanese horseshoe crab	<i>Tachypleus tridentatus</i>
Malaria mosquito	<i>Anopheles gambiae</i>
North Atlantic salmon	<i>Salmo salar</i>
Northern pike	<i>Esox lucius</i>
Pacific hagfish	<i>Eptatretus stouti</i>
palmetto or hybrid bass	<i>Morone saxatilis</i> × <i>Morone chrysops</i>
Pearl danio	<i>Danio albolineatus</i>
Planarian	<i>Dugesia japonica</i>
Purple sea urchin	<i>Stongylocentrotus purpuratus</i>
Slime mold	<i>Dictyostelium discoideum</i>
Steelhead trout	<i>Oncorhynchus mykiss</i>
Striped bass	<i>Morone saxatilis</i>
Striped loach	<i>Acanthopthalmus kuhli</i>
Three-spined stickleback	<i>Gasterosteus aculeatus</i>
Tiger pufferfish	<i>Fugu rubripes</i>
Tilapia	<i>Oreochromis mossambicus</i>
White bass	<i>Morone chrysops</i>

White cloud mountainfish

Tanichthys albonubes

White perch

Morone americana

Yellow perch

Morone mississippiensis

Zebrafish

Danio rerio

CHAPTER 1. INTRODUCTION

A synopsis of adaptive immunity

It is believed that the establishment of molecular mechanisms for self recognition (i.e. intercellular adhesion and communication) was crucial for the evolutionary emergence of multicellular organisms [1], but their evolutionary success would also appear intimately linked to their capacity for non-self recognition. Another significant evolutionary transition, from jawless- to jawed-vertebrates, marks the appearance of a substantially more complex immune system, specifically an adaptive immune system [2]. A unique feature of the immune response in vertebrate, as derived mostly from the study of mammals and birds, is that if an individual survives the initial infection, the system adapts and remains prepared, typically for the rest the animal's life, to rapidly clear any subsequent infection. Several evolutionary innovations contributed to establishing this immunological "memory" [3]. Firstly, the somatic recombination of receptor genes, which occurs independently in each lymphocyte as it matures, permits generating an exorbitant number of unique receptors to select from without needing to encode them individually in the genome. Secondly, the development of mechanisms for sorting through these cells in order to select those which are specific to infectious agents and eliminate those that are specific for self. Following is a brief review of the elements that define the adaptive immune response.

The appearance of an adaptive response [2] is correlated with the simultaneous appearance of three genes central to immune recognition: rearrangement-associated genes (RAG), the variable (V) domain of the Ig superfamily (IgSF), and class I and class II major

histocompatibility complex receptors (MHC). The V(D)J recombination that produces the V domain, which mediates recognition by immunoglobulins (Ig) and T-cell receptors (TCR), can occur only through the action of the RAG protein. Afterwards, the selection and proliferation of lymphocytes expressing the recombined V domain of high affinity to non-self requires interaction with the MHC presented on the cell surface. Through this recombination and selection process, clonal lymphocyte populations emerge that can efficiently detect pathogens present intracellularly (i.e. viruses and intracellular bacteria) or encountered in the interstices of the body. For intracellular pathogens, a subset of naïve T-cells, which have never encountered antigen, are activated into cytotoxic cells when the TCR recognizes MHC I loaded with peptides produced during degradation in the cytoplasm. A different subset of naïve T-cells are activated into effector helper cells through MHC II loaded with peptides produced during phagosomal degradation. The subsequent interactions of these activated T-cells compose cellular immunity. However, these effector helper T-cells are also responsible for initiating a humoral response through activation of mature B-cells, which will eventually mature into antibody-secreting plasma cells. Therefore, the resulting contribution of RAG, V domain recombination, and MHC selection is the development of humoral and cellular mechanisms for rapidly detecting and eliminating pathogens.

The best inducers of immunity are T-dependent antigens (TD) presented by MHC (i.e. proteins), however other macromolecules present in pathogens such as polysaccharides, nucleic acids and lipids can also serve as antigens. Since MHC does not present these molecules, they are referred to as T-lymphocyte-independent (TI) antigens. However, TI antigens are poor immunogens because they do not activate cellular immunity. Characteristic of the immune response to TI antigens is the prevalence of low affinity IgM and the lack of

establishment of lymphocyte memory so important for immunity. This difference in immunogenicity between TD and TI antigens is particularly important for vaccine development since bacterial and fungal cell surfaces are covered predominantly with polysaccharides (e.g. capsules, peptidoglycan, lipopolysaccharide, β -glucan, and mannan) serving as prominent epitopes available for immune recognition [4]. Moreover, children younger than 2 years old and the elderly appear unresponsive to polysaccharide immunizations making it difficult to protect them from infections, which is especially of concern due to the current rise in antibiotic-resistant infections. Evidently, the adaptive immune system is geared toward the recognition of peptide antigens rather than saccharides. This can sometimes be overcome by conjugating saccharides to a carrier protein to induce the presentation of glycosylated peptides by the MHC. However, this does not likely reflect the events during a natural encounter with a pathogen. What is emerging is that for recognition of TI antigens, the adaptive immune system is dependent on non-immunoglobulin receptors for initiating an immune response.

The immunoglobulin fold

The Ig fold [5] is the most frequently detected protein domain in receptors associated with the adaptive immune system. Currently, Ig domains can be classified based on sequence similarity into V, C and H domain families, which together form the Igsf, however it is the presence of hypermutable regions in the V domain that make it an extremely good receptor domain. The Ig fold is topologically described as a sandwich formed by two β -sheets consisting of 3 to 5 antiparallel β -strands connected by loops of various lengths. An intrachain disulfide bridge is commonly present serving to stabilize the fold. Substantiated by the diversity represented by

the Ig fold, the Ig fold is highly mutable in positions that do not disrupt the fold, especially the loops. It is three of these loops, named complementarity-determining regions (CDR1, 2 and 3), from the distal V domain on the immunoglobulin that form the binding site. An important property of the Ig fold is its tendency to associate to itself as is found in immunoglobulins. An immunoglobulin basically consists of two light (2L) and two heavy polypeptide chains (2H) each possessing one variable (V_L and V_H) and several constant (C_L and C_H) Ig domains. Association of L and H chains involves interactions between their C domains, but association between V domains actually results in creation of the antibodies binding site by bringing together both domains' CDRs. In other words, the binding-site of immunoglobulins is not encoded within a single polypeptide, but instead relies on the integration of CDRs from heterologous domains. Arguably, this structural feature has made the immunoglobulin fold the predominant domain of adaptive immune receptors.

A synopsis of innate immunity

An inherent deficiency of the process of acquiring immunity is the initial lag (i.e. days) required for activating lymphocytes that will eventually establish a specific response that can quickly eliminate a pathogen. During this naïve stage, the immune system depends on the general activation of humoral and cellular effector mechanisms that will typically eliminate the offending pathogen. Since the receptors involved in recognition of the pathogen are not recombined, like immunoglobulin genes, they are referred to as innate or germline-encoded. Macrophages, neutrophils and natural-killer cells, not lymphocytes, are the principal effector cells that respond to microbial invasion as part of the innate response. The complement

system, coagulation factors, collectin and pentraxins are examples of humoral effectors of innate immunity.

As mentioned, innate immunity has emerged as an important contributor during initiation of an adaptive response by recognizing TI antigens and supplying secondary signals required for activation of lymphocytes [6]. It has long been observed that an adjuvant is required to initiate an immune response toward protein antigens that are poorly immunogenic. An explanation resides in that the uptake of microbial products commonly contained in adjuvant induces the expression of accessory receptors, such as B7 receptors [7], on the surface of long lived phagocytes (i.e. dendritic cells and macrophages) during antigen presentation. T-lymphocytes are activated only if they receive this secondary signal. Similarly, B-cells also are activated through a secondary signal provided by a component of innate immunity. Circulating antigens fixed with complement component C3d are able to bind both mIg and complement receptor 2 present on the lymphocyte surface which initiates proliferation [8]. Therefore, innate recognition translates into positive signals for activation of both humoral and cellular mechanisms of the adaptive immune response.

Innate immune receptors are typically considered non-specific in comparison to antibodies because their affinities are typically lower. Nevertheless, innate receptors do specifically bind to moieties, or molecular patterns, typically present on the surface of microbes, but absent from the metazoan host. These have been referred to as pathogen-associated molecular patterns (PAMP) [9]. Examples of likely PAMPs are lipopolysaccharide (LPS), capsular polysaccharides (CPS), peptidoglycan (PG), teichoic acid (TA), lipoarabinomannan, N-formylmethionyl peptides, and unmethylated CpG dinucleotide-rich DNA from bacteria, mannose-rich cell wall polysaccharides from fungi (i.e. zymosan), and double-stranded RNA

from retroviruses. Most of these molecules appear indispensable for survival and consist of conserved repeating structures. Consequentially, the repertoire of receptors needed to recognize these structures does not probably need to reach the diversity attainable by somatic recombination of antibodies. Interestingly, most PAMPs are not polypeptides, so like most TI antigens they are not expected to be good immunogens. In the following section, a brief overview of innate receptors is given to illustrate the diversity of protein domains that are involved in recognizing PAMPs. It is noteworthy that none of the receptors currently reported possess an immunoglobulin domain.

Cellular innate receptors

Phagocytes are endowed with an array of receptors involved in the recognition of PAMPs [10]. Dendritic cells (DCs) are especially efficient at antigen uptake as they are capable of engulfing fluids (i.e. macropinocytosis), ingesting particulates (i.e. phagocytosis), and receptor-mediated endocytosis through diverse PAMP receptors such as C-type lectins and Toll-like receptors [11]. Most of the C-type lectin receptors described consist of a single lectin domain, but the macrophage receptor [12], also expressed by most myeloid cells, consists of 8 concatenated lectin domains. A more detailed discussion of the C-type and other lectin families is presented below.

Presently, much attention is directed at the Toll-like receptors (TLR) of mammals, which were initially linked to innate immunity from studies of the antimicrobial response in the fruit fly (*Drosophila melanogaster*) [13, 14]. TLRs consist of a leucine-rich repeat (LRR) ectodomain and a cytoplasmic IL-1 receptor-like domain (TIR) [15] which activate signal transduction through the NF- κ B pathway [16]. Unlike the fruit fly [17, 18], many of the mammalian TLRs appear to

directly bind, through their LRR, diverse and unrelated PAMPs [19]. However, the structural mechanism for recognition of such diversity of PAMPs remains unknown at present.

Although TLRs are not endocytic receptors, their activation initiates the adaptive response (i.e. T_H1) that is crucial for eliminating microbial infections [20] demonstrating that they also help instruct the adaptive immune response.

Acute phase proteins

Inflammation frequently accompanies the activation of an innate response. It is principally characterized by an increased blood flow to the area of injury, increased blood vessel permeability allowing plasma proteins to infiltrate, the arrival of neutrophils in great numbers to the site, and a rise in body temperature [21]. Cytokine signals that originate at the site of injury spread to the rest of the body and in the liver they induce what is referred to as an acute phase response. The liver responds to these inflammatory signals by rapidly upregulating synthesis of acute phase proteins (APP) many which are innate receptors such as pentraxins, complement proteins, and collectins. Of these APPs, pentraxins exhibit the most dramatic response by increasing as much as three orders of magnitude from their normal levels [22]. C-reactive protein (CRP), the first pentraxin identified, specifically binds to the unusual phosphorylcholine modification of teichoic acid from the pneumococcal cell wall [23]. CRP also binds lipophosphoglycan present on *Leishmania*, a protozoan parasite [24] so it appear to have broad recognition capabilities. The principal effector functions initiated by binding of CRP are phagocytosis and activation of the classical complement pathway [25].

Another acute phase protein that functions as a PAMP-receptor is the lipopolysaccharide - binding protein (LBP) [26]. As the name implies, LPB binds LPS released from Gram-negative bacteria and transfers it to CD14 [27], a GPI-anchored receptor that associates with

TLR4 [28]. Surprisingly, LBP [29] and CD14 [30] are also capable of binding to bacterial PG. Although the structure of LBP with ligand has not been determined, it is proposed that the determinants recognized by LBP are the repeating array of anionic groups (i.e. carboxyls) attached to the disaccharide repeats that form the backbone of PG. Unlike LPS, PG is present in the cell walls of all bacteria, which suggests that LBP is a sensor for bacteria in general and not only of those that are Gram-negative. It remains to be seen if upon binding PG, LBP and CD14 induce TLR4 signaling.

Lectins

Lectin generally refers to any protein, excluding recombined immunoglobulins, that binds carbohydrate, but does not exert enzymatic activity (e.g. hydrolases). Lectins were initially grouped in families that shared unique characteristics such as calcium dependence, ligand specificity, or source. With the advent of techniques for recombinant cloning of DNA, these initial representatives provided the corresponding sequence motif with which to identify homologues. The currently described lectin families are calcium-dependent C-type, galactose-binding galectins [31], immunoglobulin-like I-type [32], P-type mannose-6-phosphate receptors [33], and pentraxins [34] from animals, and legume [35] and cereal [36] lectins from plants. Presently, for each of these families a representative topological structure has been determined including placement of their binding sites as demonstrated by the co-crystallized ligand [37-42]. The diversity of animal lectins is not restricted to just these families as many other unique saccharide-binding topologies, sometimes of unclear ancestry, have been discovered [43, 44]. The functional mechanisms of some formerly well-studied proteins are being reevaluated since discovering they additionally possess lectin activity. For example, calreticulin and calnexin, named for their involvement in intracellular Ca^{2+} homeostasis, also

are involved in quality control of protein-folding within the endoplasmic reticulum, which is achieved through recognition of the polypeptide's glucosylation state [45]. Likewise, the lectin activity of certain cytokines maps to a site unique from the known peptide receptor-binding site, which led to postulating a new model for cell signal activation [46]. As direct products of genes that can be manipulated *in vitro*, protein-protein interactions have received much attention, but technological inroads [47] should facilitate understanding the contribution of less malleable post-translational modifications, such as glycosylation, to protein interactions [48]. Surely, continued interest on the role of glycoconjugates will likely reveal that lectin activity is more widespread than previously appreciated.

The diversity of lectin families suggests great divergence between them, but detailed topological comparison demonstrates they share many similarities. Some of the lectin families share similar topology, such as legume lectins, galectins and pentraxins [49], despite lacking any sequence similarity or location of binding site, suggesting a parallel evolution of saccharide-binding activity by analogous folds. Moreover, distinct topologies also appear to share similar interactions for achieving specificity towards ligand [50]. The binding sites of lectins are typically a shallow indentation where polypeptide residues extend to mediate hydrogen bonds with saccharide hydroxyls and the ring oxygen. The hydroxyls of saccharides can form cooperative hydrogen bonds by serving as both acceptor of two bonds and donor of one bond. In addition, van der Waal packing through aromatic residues that stack against the aromatic face of the monosaccharide ring plays a part in binding [51]. In the case of the C-type family of lectins [52], binding is mediated directly by the lectin-coordinated Ca^{2+} that forms coordination bonds with the hydroxyl oxygen of the saccharide. Abrogation of binding activity upon removal of the cation emphasizes the importance of this interaction. Water, as

both donor and acceptor, also plays an important role in forming hydrogen bridges especially if residues are out of reach of the saccharide hydroxyls. The most common monosaccharides present in natural glycoconjugates are mannose (Man), galactose (Gal), glucose (Glc) and their amidated derivatives (i.e. GalNAc and GlcNAc), fucose (Fuc) and sialic acid (NeuNAc). One of the key characteristic used by lectins to distinguish these saccharides is the disposition of the C-4 hydroxyl. Gal has an axial 4-OH while in mannose and glucose this moiety has an equatorial disposition. Oligosaccharides and polysaccharides, not monosaccharides are most likely the naturally relevant ligands for most lectins, so additional interactions are formed, but usually with no more than with three saccharides. An important difference between lectins and immunoglobulins is that lectins typically exhibit much lower binding affinity; nevertheless, this is typically compensated through association of many units to create a multivalent protein. Much has been gained in the understanding of the mechanisms of lectin binding, but addressing the physiological function of lectins remains a challenge, especially in light of the multiplicity of paralogues present in most organisms.

Many cell-surface lectins, especially those of the C-type family, have evolved to function in cellular immunity [53]; additionally, there is growing evidence that mammalian serum lectins are relics of an ancient humoral immune system. The ubiquity and diversity of humoral lectins, especially in taxonomic lineages, which emerged prior to the innovation of somatic recombination and clonal selection, supports the hypothesis that lectins evolutionarily preceded immunoglobulins in the role of antibodies [54]. Humoral or hemocyte-associated lectins have been detected in diverse mollusks [55, 56], arthropods [54, 57-70], non-chordate deuterostomes [71-74] and lower chordates [75] demonstrating their widespread presence [76, 77]. Study of the horseshoe crab (*Tachypleus tridentatus* and *Limulus polyphemus*), exploited for the

sensitivity of its blood coagulation system [78] to LPS [79], has provided the best description in arthropods of direct induction by a microbial product of hemocyte degranulation resulting in the release of lectins and diverse antibacterial proteins [80]. The release of whole genome sequences of the nematode (*Caenorhabditis elegans*) [81], the fruit fly [82], and the ascidian (*Ciona intestinalis*) allows the complete enumeration of the lectin gene repertoire from diverse invertebrate lineages. Evidently, multiple members of the C-type family [83-85] are present in each of these taxa, which is interesting since this domain is well known for its involvement in vertebrate immunity [86]. The role of these lectins remains to be determined, but their predominance is suggestive of their importance to invertebrate defense.

Undoubtedly, arthropods have been useful in the discovery of novel recognition mechanisms of innate immunity [87], but the lack of correspondence of invertebrate lectin genes to those in vertebrates leaves any statement of homology questionable. However, the fact that components of the lectin-activated complement pathway [88] are conserved throughout chordates [89] provides strong evidence that lectins preceded immunoglobulins as antibodies. Initial evidence of activation of a complement cascade by a lectin was observed with the rabbit mannose-binding lectin (MBL) [90], which is very similar to the prototype described from rat [91]. MBL is one of several collectins [92] sharing a common structure consisting of a filamentous collagen-like tail, a neck region, and terminated with a C-type lectin globular head. The collagenous tail assembles into an alpha-helical coiled-coil producing a trimer of lectin domains at one end [93]. In the case of MBL, these trimeric protomers further assemble into larger oligomers, which is necessary for complement activation [94, 95]. Ficolins, whose globular head is an N-acetyl amino saccharide-specific fibrinogen-like domain [96], and collectins both share the collagen-like tail of C1q [97], which is implicated in the association to

C1s and C1r serine proteases in order to form the C1 complex. This implied that like C1q, MBL and ficolins also associate with serine proteases. Subsequently, classical complement pathway activation by MBL was demonstrated to require homologous proteases, known as MBL-associated serine proteases (MASP), which were co-eluted throughout the purification of MBL [98]. Unlike C1q, MBL and ficolin bind directly to the target bacterial surfaces rather than through an Ig so they are able to initiate complement fixation directly. A survey of immune-related genes in the recently released ascidian genome [85] conclusively demonstrates that a vertebrate-like complement system was established prior to the emergence of adaptive immunity. Candidate genes for all three receptors including components of both the classical and alternative complement pathways were detected, which contrasts to the lack of any of the genes representative of adaptive immunity. This evidence suggests that C1q bound an alternative ligand, likely a pentraxin [99], prior to the appearance of Ig. Hence, lectin-activation appears to be the original mechanism for endowing specificity to complement fixation illustrating a link between ancient humoral recognition and effector mechanisms.

Immunity in ancient vertebrates

The existence of an adaptive immune response can be found throughout vertebrates except in their most primitive class, the agnathans. There is no evidence of the presence of RAG, MHC or immunoglobulin gene clusters in invertebrates though the Ig domain is employed for immune recognition [100, 101]. The relatively sudden appearance [102] of an adaptive immune response would appear an unlikely event that has been explained [103, 104] to be due to large scale chromosomal duplications, which subsequently allowed the diversification of multiple immune receptor genes. Gene duplication has been long proposed as the principal

mechanism [105] for expanding the protein repertoire [106]. It is commonly observed that for every gene from an invertebrate there are typically four homologues present in vertebrates [107]. Surprisingly, this was not just the result of small scale gene duplications, but of two virtually consecutive whole genome duplications that occurred during the transition from cephalochordates to jawed fish (600 mya) [108]. Additionally, a third duplication occurred during the radiation of modern ray-finned fish lineages (320 mya) after they diverged from the lobe-finned fish [109, 110] leading to even more gene copies than in tetrapods, although many appeared to have been subsequently lost.

Teleosts possess many of the basic features of the mammalian immune system though they do present important differences that suggest their system may be less efficient. This conclusion is principally based on comparative analysis of immunoglobulin gene structure [111] and their serum repertoire upon immunization. Firstly, affinity maturation, the phenomenon where antibody affinity increases as the immune response progresses, appears to be absent in teleosts. Secondly, electrophoretic analysis of circulating antibodies from fish reveals fewer isoforms than those detected in mammals demonstrating that their diversity is lower. In addition, fewer immunoglobulin isotypes are present in fish as only IgM-like and IgD-like isotypes have been detected in contrast with the diverse isotypes (i.e. IgM, IgG, IgA, IgE and IgD) present in mammals, which are dedicated to specific effector functions. The genetic capacity for generating diversity by means of somatic recombination and hypermutation is present in teleosts, so it may be that the deficiency lies in the selection of B-cells expressing high-affinity mIg receptors. Indeed, secondary lymphoid structures and cells associated with affinity maturation (i.e. germinal centers and follicular dendritic cells) first appeared in mammals [112]. Regardless of their lack of antibody affinity maturation, so crucial in maximizing the humoral

immune response, teleosts or ectotherms in general thrive in aquatic environments where they are continuously exposed to potential microbial pathogens. Therefore, innate receptors may have an even greater role than immunoglobulins as sensors of invasion by any microbes. This is supported by the observation that a zebrafish (*Danio rerio*) mutant deficient in immunoglobulin gene recombination (i.e. *rag1*) does not appear to be immuno-compromised [113].

By definition, innate receptors do not somatically recombine like immunoglobulins in vertebrates, so it is unclear then how they can express the diversity of specificities in lieu of the inefficiency of acquired immunity in teleosts. Evidence is being provided showing that teleosts possess many more members of innate receptor families in comparison to mammals. For example, the repertoire of TLRs detected in teleosts is more numerous [114-116] than that of mammals [19]. This implies that in teleosts the repertoire of TLR microbial ligands may be even more diverse than presently described in mammals. Humoral receptors also appear to be more diversified in teleosts. A single C3 is the central component of the non-specific alternative complement pathway in mammals, but in fish several isoforms are present that are specific towards diverse surfaces [117]. These examples clearly illustrate that innate receptors are diversified in teleosts potentially fulfilling functions analogous to that of antibodies. Like mammals, fish possess many of the same innate humoral receptors implicated in pathogen recognition, such as collectins [118], pentraxins [119-122], and LPS-binding protein [123]. However, unlike TLRs, it does not appear that these receptors are more diversified than their mammalian homologues.

Components of the lectin-activated complement pathway have been identified in teleosts [124, 125], but the additional presence of diverse humoral lectins [126-129] suggests that lectins may

have a more prominent role in humoral immunity of teleosts than in mammals. Like collectins, these lectins may participate in complement activation, but likely perform as immune effectors through agglutination, neutralization or opsonization.

Fish as models of immunity

Significant advances have been made to develop teleostean models for the study of immunity especially the establishment of the tools comparable to those that are available for mice and humans. Immortalized lymphocyte and macrophage cell lines finally have been established [130, 131], which are crucial for studying cellular immunity. The number of cytokines identified and characterized is increasing [132] which should provide useful reagents for modulating the immune response. As previously mentioned, teleosts have undergone a genome duplication that may have allowed their protein repertoire to expand extensively. The diversity of the C3 complement components may have resulted from this event. There have even been advances in developing zebrafish, a popular development model, as a model for immunology [133]. Surely, teleost are increasingly contributing insights into both ancient and derived features of vertebrate immunity.

Project Goal and Significance

The goal of this dissertation project is the identification and characterization of humoral lectins from a ray-finned fish and to describe the role they may serve as innate PAMP receptors. Insight into conserved mechanisms of innate immunity may be gained from the study of organisms not conventionally perceived as models for the study of human immunology. Specifically, ray-finned fish represent a transitional stage towards the development of the

mammalian adaptive immune system since they exhibit lower capacity for generating diversity of recombined immunoglobulin-receptors. Therefore, ray-finned fish potentially rely substantially on humoral innate receptors, such as lectins, for immediate recognition and elimination of potential pathogens thereby compensating for their underdeveloped antibody repertoire. In this immunological context, the contribution of lectins in defense may be more readily parsed whereas in mammals efficient development of immunological memory may occlude elucidating their function. A basic understanding of innate immunity in teleosts is not only potentially valuable from a comparative standpoint, but can also contribute to successful management of infectious diseases in cultured fish stocks by improving vaccine development [134].

CHAPTER 2. IDENTIFICATION OF A UNIQUE BASS LECTIN

Moronid bass are commonly found inhabiting fresh and saline waters of the North Temperate Zone. The four North American species currently described are striped bass (*M. saxatilis*), white bass (*M. chrysops*), white perch (*M. americana*) and yellow bass (*M. mississippiensis*). Species phylogeny based on morphological and molecular characters suggests these species group into two sister clades: white perch: yellow bass and striped bass: white bass, the former being the most closely related [135]. The anadromous life style of the striped bass is unique from the other species, which exclusively inhabit fresh water. Rockfish, as striped bass is also known, are the largest growing within the genus and due to this, meat quality, and angling characteristics they have been historically of commercial importance [136]. Until the mid 1800s, the striped bass' grounds stretched only from the Gulf of Mexico to the Maine coast. Populations present along the Pacific coastline appeared later on due to introduction from humans. In the 1960s, a successful hybridization of the striped bass (♀) to the white bass (♂) created the hybrid palmetto bass [137]. The motive behind creating hybrid crosses was to combine the size, longevity, food habits, and angling qualities of the striped bass with the adaptability to exotic environments of the white bass. Indeed, the palmetto bass exhibits heterosis showing improved survivability, superior early growth rates, greater disease resistance [138, 139], and general hardiness [140] compared to striped bass. These features have been exploited for aquaculture by allowing hybrids to be raised at higher population densities required to maintain efficient production. The indoor culture of palmetto bass in the mid-atlantic region presently allows substantial numbers of large fish to be readily obtained, which

was especially advantageous for obtaining the biological materials required for this research project.

At the onset of this project, a choice was made between following a molecular or a biochemical approach. Identification of lectins through use of PCR primers based on protein family profiles can be an uncertain proposition [141] due to the extensive divergence observed in their polypeptide sequence. Therefore, a purification protocol exploiting lectin activity and documented dependability of sugar-affinity chromatography was chosen. In addition, carbohydrate-affinity purification targets the defining activity of lectins rather than focusing on any single lectin sequence family. From the apparent ubiquity of MPB-like proteins in vertebrates, one lectin purification protocol was designed based on the successful methodology implemented for isolation of collectins from the liver. As a first step in identifying a collectin-like lectin from the palmetto bass liver, a protein purification protocol was compiled based on the published methods for the rabbit [142], rat [143], alligator [144], and human [145] liver collectins. These protocols center on the use of mannan-affinity chromatography to selectively exploit the carbohydrate-binding activity commonly exhibited by collectins. However, to cover the possibility that the bass possesses lectins of alternate specificity, other ligands (i.e. Fuc, Gal, GlcNAc, Glc, and Lac) were also conjugated to agarose for affinity chromatography. Following the published methods, the initial step in purification involves extracting the liver lectin under high salt and in the presence of detergent to help dissociate the protein from a liver acetone cake. The extract is subsequently screened on each of the monosaccharide columns, which are eluted with the corresponding free monosaccharide. One drawback brought about by the use of high salt concentrations and detergent is that they are not compatible with the hemagglutination assay commonly used for

detecting lectin activity leaving no other facile means of testing for a lectin in the liver. Therefore, an alternative protocol adapted from [146] based on aqueous extraction at physiological salinity and without detergent was implemented. Based on the common presence of soluble lectins in circulation among both protostome, deuterostome invertebrates and vertebrates, as previously mentioned, blood from bass was also assayed for lectin activity. Serum was used instead of plasma since by being defibrinated it is more stable under storage and handling. This chapter describes the purification, characterization, and cloning of a fucose-binding lectin present in the liver and blood of the palmetto bass and striped bass.

Materials and Methods

Reagents

Monosaccharides, oligosaccharides, glycoproteins, divinyl sulfone, and Sepharose 6B were purchased from Sigma (St. Louis, MO, USA). Enzymes used for molecular biology were purchased from New England Biolabs (Beverly, MA, USA). Red blood cells were obtained as expired lots from the University of Maryland at Baltimore Hospital Blood Bank, or as washed Immucor Referencells (Norcross, GA, USA). EnzyOne 2000 DNA polymerase and dNTPs used in PCR were purchased from ID Labs (London, Ontario, Canada). TaKaRa ExTaq high fidelity DNA polymerase was purchased from PanVera (Madison, WI, USA).

Animals

Striped and palmetto bass of >1 kg housed at the COMB Aquaculture Research Center (ARC) were used for extraction of blood and liver. Animals for the challenge experiments (Sixteen-month-old palmetto bass; 300-400 g each) were purchased from Integrated Food

Technologies (Emmaus, PA, USA), transported to the ARC, and acclimated (10 ppt at 18 °C). Fish were held in 6 ft and 12 ft circular tanks within a recirculating water system, with stocking densities not exceeding 38 kg/m³. Feed was pelleted fish feed (Purina Mills; Richmond, IN, USA). Before venipuncture or challenge injections, all animals were anesthetized in 250 ppm 2-phenoxyethanol (J. T. Baker; Phillipsburg, NJ, USA). Killing of fish prior to dissection was accomplished by anaesthetic overdose (1000 ppm). The Institutional Animal Care and Use Committee of the University of Maryland Biotechnology Institute approved the procedures described below.

Extraction of blood and separation of serum

Peripheral blood (approx. 10 ml/kg) was collected from the caudal vein using 20 ml syringe fitted with an 18-gauge hypodermic needle (BD Biosciences; Bedford, MA, USA). After removing the needle from the syringe, the blood was transferred to 50 ml polypropylene conical tubes (Sarsted; Newton, NC, USA) and the clot allowed to retract overnight at 4 °C. Serum was separated from the clot by centrifugation at 3,000 x g for 30 min in a swinging-bucket Beckman Model J centrifuge at 4 °C, and stored at -80 °C.

Dissection and collection of livers

Whole livers (approximately 200 g each) were collected and snap frozen in liquid nitrogen. Livers were subsequently stored at -20 °C (for protein purification) and at -80 °C (for RNA extraction) until further use.

Preparation of liver extracts

Livers were first allowed to thaw overnight at 4 °C and then minced with scissors in a tray on wet ice. For extraction from acetone cake preparations, approximately 150 g minced liver was

homogenized in cold acetone (-20 °C; 2 ml acetone/g liver) for 1 min with a Waring blender to produce a fine slurry. The homogenate was filtered through Whatman No. 1 paper on a Büchner funnel and the retentate was extracted again by pouring cold acetone over the powder remaining on the filter. The resulting acetone cake was dried overnight in a vacuum chamber at 4 °C. Aqueous extraction of the delipidated liver tissue was a modification of the procedure established for purifying the alligator hepatic binding protein [144]. The acetone cake was rehydrated by suspending the powder in 500 ml (3.5 vol) of ice chilled 0.2 M NaCl, and stirring the slurry for 30 min at 4 °C. The rehydrated acetone powder was centrifuged at 12,000 $\times g$ for 15 min at 4 °C in a Beckman Model JR-21 centrifuge with a Beckman J-10 rotor. The resulting pellet was resuspended in 400 ml (2.5 vol.) extraction buffer (10 mM Tris-HCl (pH 7.8, 4 °C), 0.4 M KCl, 2% (v/v) Triton X-100, 0.1 mM PMSF, 5 mM iodoacetamide) and stirred for 30 min at 4 °C. The resulting extract was clarified by centrifugation at 12,000 $\times g$ for 15 min at 4 °C and the supernatant was set aside. The remaining pellet was extracted three more times, supernatants pooled (1.2 liters), and the final pellet discarded. Finally, the pooled extract was adjusted to 20 mM Ca^{2+} by adding solid CaCl_2 .

For preparation of aqueous liver extracts, minced bass liver was mixed with 200 ml of chilled extraction buffer (10 mM Tris-HCl (pH 7.8, 4 °C), 100 mM lactose, 25 mM KCl, 0.1 mM PMSF, 2 $\mu\text{g}/\text{ml}$ aprotinin, 1 $\mu\text{g}/\text{ml}$ leupeptin, 1 $\mu\text{g}/\text{ml}$ pepstatin), and liquefied in a chilled Waring blender for 1 min. The smooth homogenate was centrifuged at 12,000 $\times g$, 15 min at 4 °C. The supernatant was filtered through cheesecloth, and stored on wet ice. A second extraction of the pellet with 200 ml chilled extraction buffer was performed twice, and the supernatants pooled with the first supernatant. To allow lipids to separate, the homogenate

was kept on wet ice for 1 h, and the floating lipid layer removed with a pipet. The homogenate was clarified by a second centrifugation at $14,000 \times g$ for 1 h at 4 °C.

Protein concentration estimation

Protein concentrations were estimated using the Protein Assay kit I (BioRad; Hercules, CA, USA), based on Coomassie dye binding [147] and crystalline bovine serum albumin as standard. The microassay format in 96-well flat bottom plates was performed as described by the manufacturer. Five minutes after combining protein and dye, the reactions were read at 595 nm on a SpectraMax 340 plate reader (Molecular Devices; Sunnyvale, CA, USA) controlled by SoftmaxPro software v.1 (PC version).

Protein purification and storage

A. Preparation of affinity chromatography gel

Sugars were covalently bound to the chromatography resin by the divinylsulfone method [148], which has the advantage of producing 40-50 μmol active groups per ml of gel and therefore produces a highly substituted matrix. This method involves the reaction of the sulfone-bridged divinyl group to the hydroxyl-rich chromatography gel and subsequently forming a glycosidic linkage to the free saccharide. The protocol is as follows: 100 ml settled volume of Sepharose 6B-CL (Sigma, St. Louis, MO, USA) was washed with 1 l water in a medium porosity sintered glass funnel, suction-dried to a wet cake, and transferred to a 500 ml glass beaker. The gel was suspended in 100 ml 0.5 M sodium carbonate and mixed with magnetic stirrer. In a flow hood, a volume of 10 ml divinylsulfone (DVS) (Fluka; Milwaukee, WI, USA) was added drop-wise with constant stirring over a 15 min period after which the suspension was continually stirred for 1 hr at room temperature. After this activation step, the gel was

washed with 2 l of water in the sintered funnel until the filtrate was no longer alkaline. To couple the carbohydrate ligands, the activated gel was washed with 0.5 M sodium carbonate (5 vol.) on the sintered funnel and suspended in an equal volume of a 20% solution of desired saccharide dissolved in 0.5 M sodium carbonate. Man, GlcNAc, Fuc, Lac and Glc were conjugated in separate batches. During coupling the gel was mixed end-over-end at room temperature for 24 hr. After the coupling period the gel was washed successively with 2 l of water and 2 l of 0.5 M sodium bicarbonate on the sintered funnel. To block any unreacted vinyl groups present, the conjugated gel was suspended in 100 ml 0.5 M sodium bicarbonate containing 5 ml 2- β -mercaptoethanol (2-ME)(BioRad; Hercules, CA, USA), and stirred for 2 hr at room temperature. Finally, the gel was washed with 2 l of distilled water and stored in the presence of 0.02% (w/v) sodium azide at 4 °C to inhibit bacterial growth.

B. Affinity chromatography of liver acetone extracts

Liver extracts were applied (0.9 ml/h flow rate) to a 60 ml bed volume of sugar-conjugated Sepharose 6B-CL pre-equilibrated with equilibration buffer (10 mM Tris-HCl, pH 7.8, 4 °C; 20 mM CaCl₂; 1.25 M NaCl; 0.5% (v/v) Triton X-100) in a 2.5 cm i.d. Kontes Flexcolumn (Vineland, NJ, USA). The low pressure chromatography system consisted of a Rainin Rabbit Peristaltic Pump (Woburn, MA, USA), a LKB 2112 Redirac Fraction collector (Upsala, Sweden) and a LKB Uvicord S ultraviolet detector placed in a walk-in refrigerator. All affinity chromatography steps were performed at 4 °C. After washing the column to baseline absorbance as monitored at 280 nm, the potentially-bound fraction was eluted from the column with 200 mM of the corresponding sugar ligand dissolved in equilibration buffer or eluted with EDTA elution buffer (20 mM Tris-HCl, pH 7.8, 4 °C; 1.25 M NaCl; 20 mM EDTA; 0.5% (v/v) Triton X-100). The eluate was dialyzed using a SpectraPor 1 membrane

(Spectrum Medical Supplier; Carson, CA, USA) against 0.05% (v/v) Triton X-100, and the detergent removed by protein precipitation in cold ethanol [149]. An equal volume of cold ethanol (-80 °C) was added and the mixture left standing for 10 min in ice water. The precipitated protein was pelleted by centrifuging at 20,000 $\times g$ for 10 min, at 4 °C in a Beckman J-20 rotor. Finally, the pellet was washed in 50% (v/v) ethanol and resuspended in deionized water or buffer.

C. Anion exchange batch chromatography (AEX) of aqueous liver extract

To address the possibility that lectin-bound endogenous ligands may interfere with the affinity chromatography isolation, batch anion-exchange chromatography was performed prior to the affinity step. The pH of the aqueous liver extract prepared in the previous step was adjusted to 8.5 with 1 M Tris and combined with pre-equilibrated 100 ml packed DEAE-Sepharose-4B (Sigma) by mixing on an orbital shaker for 2 h at 4 °C. The chromatography gel was washed on a sintered glass funnel with 10 volumes of AEX buffer (10 mM Tris-HCl, pH 8; 25 mM KCl; 0.1 mM PMSF) to remove any putative endogenous ligand of the lectin. Protein was eluted by step elution in 0.25 M, 0.5 M and 1 M NaCl in 10 mM Tris-HCl, pH 7.8, 4 °C; 0.1 mM PMSF.

D. Affinity chromatography of aqueous liver extracts

The subsequent step after AEX was affinity chromatography on Glc, Man, GlcNAc, Lac or Fuc-conjugated Sepharose 6B-CL packed in a 1.5 cm i.d. Kontes Flexcolumn. The packed column was preequilibrated in 10 mM Tris-HCl, pH 7.8, 4 °C; 20 mM CaCl₂; 0.5 M NaCl; 0.1 mM PMSF. After adjusting to 20 mM Ca²⁺ with solid CaCl₂, the lysate was passed at 0.9 ml/h and the flow monitored at 280 nm. The column was washed to baseline absorbance with equilibration buffer, and the lectin(s) bound in each column eluted with 200 mM of the

corresponding ligand sugar in equilibration buffer. The eluates were dialyzed against 50 mM Tris-HCl, pH 7.5, 4 °C; 100 mM NaCl; 20 mM CaCl₂ to remove the eluting sugar, and stored at 4 °C.

E. Affinity chromatography of serum

Typically, serum samples from four specimens were pooled (40 ml), dialyzed against 2 l of TBS-Ca (50 mM Tris-HCl, pH 7.5, 4 °C; 100 mM NaCl; 20 mM CaCl₂) with four changes in a 24 h period. The dialyzed serum was diluted 1:1 in TBS-Ca and passed over a 10 ml bed volume of saccharide-linked gel (2.5 cm i.d. Kontes Flexcolumn) at 0.9 ml/h and monitored at 280 nm. The column was washed to baseline with TBS-Ca, eluted with 30 mM EDTA in TBS (no CaCl₂), washed to baseline again, and finally, eluted with 200 mM Fuc in TBS-Ca. Fractions (0.5 ml) containing protein peaks were pooled and concentrated by ultrafiltration on a Centricon 10 (Millipore; Bedford, MA, USA) to a final concentration of approximately 10 mg/ml. An opportunity for purifying of gram quantities of protein presented itself when significant amounts of blood (~100 ml/fish) became available due to the need for euthanizing diseased large striped bass (>3 kg).

F. Size exclusion chromatography

The concentrated Fuc-binding fraction from the affinity step was polished by size exclusion chromatography (SEC) on a Superose 12 10/30 HR column (Amersham Biosciences; Piscataway, NJ, USA) pre-equilibrated with 0.45 µm-filtered running buffer (TBS-Ca, 10 mM lactose, 0.02% sodium azide). This was performed using a Beckman Gold 126/166 HPLC system (Fullerton, CA, USA) at a flow rate of 0.4 ml/min, at room temperature, and the elution monitored at 280 nm through a 1 cm path-length cell. Fractions were collected by hand into polypropylene Eppendorf microtubes.

G. Storage of purified bass lectin

For short term storage, usually no more than 1 month, 1 mg of purified protein was stored at 4 °C in the presence of 0.02% (w/v) sodium azide and the rest frozen at -20 °C for long term storage.

Analytical procedures

A. Electrophoresis and isoelectric focusing

Polyacrylamide gel electrophoresis in the presence of dodecyl sodium sulfate (SDS-PAGE) [150] was performed on 5-20%T gradient mini-gels (8x10 cm) using a Hoefer SE250 system (Amersham Biosciences) cooled at 15 °C. The gels were prepared with reagents purchased from BioRad (Hercules, CA, USA), using the Hoefer casting mold and the Hoefer gradient maker, following the manufacturer's protocol. Gels were stained for ≥ 2 hours in a Coomassie stain (0.25% (w/v) Coomassie R-250 in 50% methanol, 10% (v/v) acetic acid) and destained overnight in 10% (v/v) methanol, 7.5% (v/v) acetic acid with a crumpled Kimwipe to adsorb the dye. Silver staining was performed using the Silver Stain Kit (Pierce; Rockford, IL, USA) following the manufacturer's protocol. Two-dimensional gel electrophoresis was performed according to the manufacturer using the same Hoefer SE250 mini-format system equipped with a Hoefer Mighty Small™ tube gel adapter for the 1D (IEF) separation step. Urea vertical slab isoelectric focusing was performed as described in [151] using a Hoefer SE600 device.

B. Glycosylation analysis

Glycosylation was analyzed with the Glycan Detection Kit (Roche Molecular Biochemicals; Indianapolis, IN, USA), which employs an immunoblot format. Briefly, the bass lectin was resolved by SDS-PAGE and electrotransferred for 1 h at 0.4 mA in Towbin buffer (25 mM Tris, 192 mM glycine, 10% methanol) onto an Immobilon-P membrane (Millipore; Bedford,

MA, USA) using a Hoefer TE22 tank transfer unit refrigerated to 10 °C. After blocking with the solution provided by the kit, the membrane was treated with 10 mM sodium metaperiodate. This treatment cleaves vicinal hydroxyls on sugar rings, which can react with digoxigenin-succinyl-[ε]-aminocaproic acid hydrazide, an epitope tag. The labeled-lectin was incubated with a 1:1000 dilution of goat anti-digoxigenin antibody labeled with alkaline phosphatase. The bound antibody was detected by developing the membrane in the colorimetric substrate, nitroblue tetrazolium/5-bromo-4-chloro-3-indolyl phosphate (NBT/BCIP). Transferrin and creatinase served as positive and negative controls, respectively.

C. Electron spray ionization-mass spectrometry (ESI-MS)

Mass spectrometry was performed at Emory University's Winship Cancer Center Microchemical Facility. Bass serum lectin (1 nmoles) was desalted by binding to a Millipore C-18 ZipTip followed by elution with 2 µl of 50% acetonitrile: 0.1% trifluoroacetic acid. Isopropanol:water (50:50; 5 µl) mixture was added to the eluate, and 5 µl taken for flow injection analysis (FLA) on a Applied Biosystems PE-SCIEX API 3000 triple quadrupole mass spectrometer (Foster City, CA, USA). A Q1 scan was performed with a scan range of 400 to 2000 m/z, step size of 0.2 amu, and a dwell time of 0.8 ms.

D. Size exclusion chromatography

To determine the native molecular size of the Fuc-binding lectin in solution, the Superose 12 10/30 HR column used above for final purification was calibrated with low-molecular size SEC standards of known hydrodynamic radius (Amersham Biosciences) following the procedure in [151] while using the same HPLC system. These calibration markers were: bovine serum albumin (BSA; 35.5 Å, 67 kDa), ovalbumin (30.5 Å, 43 kDa), chymotrypsinogen

A (20.9 Å, 25 kDa), and RNase A (16.4 Å, 13.7 kDa). Two sets of calibration protein mixes were prepared in water at empirically determined concentrations to yield peaks of similar absorbance at 280 nm. Calibration marker set I consisted of BSA (7 mg/ml) and chymotrypsinogen A (3 mg/ml). Calibration marker set II consisted of ovalbumin (7 mg/ml) and RNase A (10 mg/ml). Injecting 200 µl of each set separately allowed complete peak separation for each of the marker proteins using this specific column. The column void volume (V_0) and total liquid column volume (V_t) was determined by injecting 200 µl of dextran blue 2000 (2 mg/ml) (Amersham Biosciences) dissolved in a solution of 5 mg/ml chromatography-grade acetone (0.792 g/cm^3) (Fisher Scientific) in water. For the purpose of normalizing the procedure irrespective of chromatography system, a distribution coefficient (K_D) was calculated as follows: $K_D = (V_r - V_0)/(V_t - V_0)$, where V_r is the elution volume of a peak. A plot of K_D vs \log_{10} MW was prepared and a third order polynomial curve fit was made (DeltaGraph v.4.1, SPSS; Chicago, IL, USA) to allow the interpolation of molecular weight for the experimental sample. The estimate of molecular weight assumes that the experimental protein is globular and behaves hydrodynamically similar to the calibration markers.

E. Chemical crosslinking

MsaFBP32 was first dialyzed against 2 l of 100 mM sodium HEPES (pH 7.5), 150 mM NaCl, 10 mM CaCl_2 . Concentration of working solution was adjusted to 35 µM (1.13 g/L based on MW 32,388 Da) and 5 µl aliquots were combined with 5 µl (0, 35, 70, 140, 350, 700, 1,400 µM) bis(sulfosuccinimidyl) suberate (BS3, Pierce Chemicals) (10 mM stock and serial dilutions were prepared in fresh dialysis buffer) to achieve up to a 20-molar excess. The reactions were incubated for 60 min at room temperature and finally diluted 1:1 with double-strength SDS-PAGE (DTT) sample buffer, heated to 100 °C for 5 min to denature, and the full reaction (20

μl) loaded on a 10% SDS-polyacrylamide mini-format gel (1.5 mm thick) (Hoefer SE250).

Molecular weight calibration markers were obtained from BioRad (Broad SDS-PAGE).

Carbohydrate-binding properties

A. Hemagglutination assay

Erythrocytes (red blood cells, RBCs) were used in an agglutination tests after Pronase treatment (Pr-RBC) using 96-well Terasaki plates (Robbins Scientific; Mountain View, CA, USA) as reported earlier [152]. Plates were blocked overnight at 4 °C in 1% (w/v) BSA (Sigma) in physiological saline (0.85% NaCl, 0.2% sodium azide). Dilution plates (96-well, Corning; Acton, MA, USA) were blocked to avoid loss of lectin during dilution. The RBCs were washed 4 times at 800 x g for 2 min in physiological saline, and 50 μl of packed RBCs were resuspended in 50 μl of 0.1% Pronase (Calbiochem-Novabiochem; La Jolla, CA, USA) in physiological saline for 20 min at 37 °C. Subsequently, cells were washed 4 times with physiological saline, and once with TBS-Ca (50 mM Tris-HCl (pH 7.6), 100 mM NaCl, 10 mM CaCl₂). The Pr-RBC were resuspended in assay buffer, counted with a hemocytometer, and resuspended at 5x10⁶ cells/ml in assay buffer. An equal volume of a Pr-RBC suspension was added to 5 μl of two-fold dilutions of bass lectin in TBS-Ca. Negative controls were set up using TBS-Ca instead of lectin. The plates were gently vortex-mixed for 10 sec and incubated at room temperature for 1 h. Finally, agglutination was assessed under a microscope and scored from 0 (negative) to +4 (full agglutination). The reciprocal of the highest dilution of lectin showing an agglutination score of +1/2 was recorded as the titer. Total agglutination units (AU) is defined as the product of the total volume of the solution assayed for

agglutinating activity and its titer. The lectin specific activity (SA) is defined as the ratio of AU per the total protein in the assayed solution.

B. Hemagglutination inhibition

The carbohydrate ligand specificity was examined by the microhemagglutination assay described above. A lectin solution with a titer of 2 AU was prepared in TBS-Ca. Mucins were desialylated by mild hydrolysis in 0.1 N H₂SO₄ at 80 °C for 1 h. All carbohydrates to be tested as inhibitors were dissolved in TBS at concentrations up to 100 mM for mono- and oligosaccharides and 10 mg/ml for polysaccharides and glycoproteins, and adjusted to pH 7.6 with concentrated NaOH. Serial two-fold dilutions were made of the inhibitors in the same buffer. Inhibition by mono- and oligosaccharides was tested at concentrations from 0.1 mM up to 200 mM. Inhibition by glycoproteins, asialoglycoproteins, and polysaccharides was tested at concentrations up to 10 mg/ml. A volume of 5 µl of ligand dilution was added to 5 µl of lectin in the wells of 96-well Terasaki plates. The mixture was incubated for 45 min and 2 µl of the Pr-RBC suspension (10⁷ Pr-RBC/ml) was added to each well. Controls are the substitutions of the inhibitor solution by TBS-Ca and substitution of purified lectins by TBS-Ca. Ligand concentrations that produced 50% inhibition (IC₅₀) were interpolated by polynomial curve fit from plots of percent inhibition versus inhibitor concentration.

C. Cation requirements

To test cation requirements for lectin binding to ligand, the bass lectin (200 µl) was dialyzed overnight against 1 l of 100 mM citrate (pH 6); 10 mM EDTA using three buffer changes. The sample was then dialyzed against TBS-HCl (pH 7.5, 4 °C) to restore neutral pH. A control sample was dialyzed against TBS-HCl (pH 7.5, 4 °C); 1mM CaCl₂ in parallel with the experimental. To test if calcium concentration has an enhancing effect on hemagglutinating

activity, 5 μ l samples of bass lectin serially diluted in TBS (No Ca^{2+}) were incubated with 5 μ l of 200, 20, 2, 0.2 and 0.02 CaCl_2 or 200 mM EDTA (pH 8) in Terasaki plates. After 30 min equilibration, 2 μ l of Pr-RBC suspension (10^7 cells/ml) were added to the wells and vortex mixed gently for 10 sec. Agglutination was read as described above.

Peptide and nucleic acid sequence

A. Edman sequence analysis

Peptide sequencing was performed at Winship Cancer Center's Microchemical Facility of Emory University. Bass lectin (50 μ g) was digested with trypsin and lysyl endopeptidase in 0.05 M Tris-HCl (pH 8.5), 1 M guanidine hydrochloride (E/S=1:50, w/w, at 30 °C, 10-20 h). The digests were acidified to pH~2 using 10% TFA, and the fragments were purified by HPLC. The peptides were separated on a microbore RP-HPLC system consisting of Applied Biosystems model 140A pumps and model 1000S diode-array detector (2.3 ml flow cell, 0.0025 inch i.d. tubing) (Foster City, CA, USA). Fractionation of the peptides was performed either on a Zorbax-SB C-18 silica column (0.1x15 cm, d_p ~5 mm, 300 Å pore size; Microtech Scientific; Saratoga, CA, USA), or on an Applied Biosystems Aquapore ODS-300 C-18 silica column (0.1 x 25 cm, d_p ~7 mm, 300 Å pore size) equilibrated at room temperature in 0.1% aqueous TFA, and eluted at flow rates of 50-80 μ l/min using linear gradients of acetonitrile/water/TFA (80:20:0.1, v/v). The column effluent was monitored at 215 nm, the UV absorption spectra of the absorbing material was determined, and the eluate was manually collected and stored at -20 °C prior to further analysis. Rechromatography of some fractions in the second RP-HPLC elution solvent system consisting of 2-propanol/acetonitrile/water/TFA (70:20:10:0.1, v/v) was required prior to sequencing.

Automated Edman degradation of the peptides [153] was performed on an Applied Biosystems model 477A/120A pulsed-liquid sequencing system (Foster City, CA, USA). The PTH-amino acids are separated and identified as described in [154].

NH₂-terminus sequencing of 10 µg of palmetto bass serum lectin was performed at the Bioanalytical laboratory of the Center of Marine Biotechnology, Baltimore, Maryland by Dr. Michael S. Quesenberry. A Beckman LG3000 gas-phase sequencer was used for this procedure. Seven cycles were completed successfully providing sequence of seven residues.

B. Isolation of total RNA from bass liver

Total RNA from liver was purified by the acid phenol/guanidinium isothiocyanate method [155] with modifications. One gram of liver tissue, stored at -80 °C, was placed in a mortar pre-chilled on dry ice and filled with liquid N₂. The tissue was ground with the pestle to a fine powder under liquid N₂, was and placed in a 50 ml polypropylene conical tube prechilled with liquid N₂. The nitrogen was allowed to evaporate and 10 ml of lysis solution (4 M guanidinium thiocyanate; 25 mM sodium citrate; 0.5% sarkosyl, 100 mM 2-ME) were added. The suspension was quickly processed with a homogenizer (Pro Scientific; Oxford, CT, USA) until complete dissolution of the tissue. The lysate was transferred to a polypropylene 30 ml Oak Ridge centrifuge tube, 1 ml of 2 M sodium acetate (pH 4) was added, and the contents mixed by vortexing. To extract DNA and protein, 10 ml of water-saturated phenol was added and mixed by vortex. A volume of 2 ml of chloroform/isoamyl alcohol (49:1) was finally added, mixed, and left on ice for 15 min. Separation of phases was accomplished by centrifugation 12,000 × g (4 °C) for 10 min in a fixed-angle rotor. To remove unwanted polysaccharides [156] from the aqueous phase collected, 5 ml of high-salt precipitation solution (1.2 M NaCl; 0.8 M sodium citrate) was added, mixed, and RNA precipitated by addition of 5 ml of isopropanol

and mixing. The precipitated RNA was collected by centrifugation at $12,000 \times g$ (4°C) for 20 min. After washing the pellet in 70% (v/v) ethanol, the pellet was dissolved in 5 ml of Tris-HCl (pH 7.5), 1 mM EDTA, 0.5% (w/v) SDS. The SDS was extracted with 5 ml of chloroform, followed by centrifugation. The RNA was precipitated by addition of 0.5 ml 3 M sodium acetate (pH 5.2) and 5 ml isopropanol, mixing between additions. To increase recoveries, the RNA was placed at -80°C for 1 h. RNA was finally pelleted by centrifugation at $12,000 \times g$ (4°C) for 30 min. The pellet was washed with 70% (v/v) ethanol, air-dried until transparent, and dissolved in 1 mM sodium citrate (pH 6.4). RNA was quantified by photospectrometry at 260 nm (1 O.D. = $40 \mu\text{g/ml}$), and the purity assessed by the ratio of 260 nm/280 nm of >1.8 O.D [157]. The integrity of the RNA was revealed by sharp bands at 2 and 5 kb (i.e. nuclear ribosomal RNA) after electrophoresis in a 1% (w/v) agarose submarine gel.

The purification of poly(A)⁺ RNA followed a conventional method based on oligo(dT)-cellulose affinity chromatography [158]. Briefly, the total RNA was denatured by heating 65°C for 5 min, cooled, on ice, mixed with 1 volume of 2X column-loading buffer (40 mM Tris-HCl, pH 7.6; 1 M NaCl; 2 mM EDTA; 0.2% (w/v) sarkosyl), and applied to a column packed with oligo(dT)-cellulose. The column was washed with 10 column volumes of 1X column-loading buffer, and the poly(A)⁺ RNA eluted with 3 column volumes of elution buffer (10 mM Tris-HCl (pH 7.6), 1 mM EDTA, 0.05% SDS), with fractions of 1/3 to 1/2 of the column volume collected. A second round of purification to remove the remaining ribosomal RNA was carried out as follows: the RNA-containing fractions from the first round of purification were adjusted to 0.5 M NaCl from a 5 M stock, heat-denatured, and chromatographed as before. The re-purified poly(A)⁺ RNA was precipitated by addition

1/10th the eluted volume of 3 M sodium acetate (pH 5.2), followed by addition of 2.5 volumes of ice-cold ethanol, and storage overnight at -20 °C. The precipitated poly(A)⁺ RNA was pelleted by centrifugation at 10,000 x *g* for 15 minutes at 4 °C, washed with 70% (v/v) ethanol, to remove salts, and centrifuged. After air-drying, the pellet was dissolved in 1 mM sodium citrate (pH 6.4). For long term storage, the poly(A)⁺ RNA was ethanol-precipitated as above, and stored at -80 °C under ethanol.

C. Synthesis of first strand cDNA

Reverse transcription of liver total RNA was performed using the SUPERScript First-Strand Synthesis System (Invitrogen; Carlsbad, CA, USA) according to the manufacturer's instructions. Five µg of total RNA was suspended in 10 µl DEPC-treated water and 1 µl (0.5 µg) of oligo dT₁₇ was added along with 1 µl of 10 mM dNTPs in an RNase-free 0.5 ml Eppendorf tube. To disrupt any RNA secondary structure, the sample was heated at 65 °C for 5 min on a dry heating block, and immediately cooled on wet ice for 1 min. After centrifugation at 14,000 x *g* for 5 s, 4 µl of RT buffer (5X), 1 µl RNaseOUT inhibitor (40 U/µl), and 2 µl 0.1M DTT were added to the sample kept on ice, and mixed by pipetting. Superscript II reverse transcriptase (1 µl of 50 U/µl) was added to the sample and was mixed by pipeting. The reaction was catalyzed at 42 °C for 50 min in an air incubator to avoid evaporation and changes in volume due to condensation on the cap. To stop the reaction it was heated at 70 °C for 15 min to denature the enzyme. Treatment with RNase H was recommended to remove the remaining RNA. After cooling the sample on ice, 1 µl of RNase H (2 U/µl) was added and the first-strand cDNA heated at 37 °C for 20 min. Finally, the cDNA was stored at -20 °C until needed.

D. cDNA cloning

The lectin cDNA was amplified by PCR by using mixed oligonucleotide amplification of cDNA (MOPAC) [159]. Degenerate primers, with 5' adapter extension ends for ease of cloning, were designed based on the NH₂-terminal and internal peptide sequence of hybrid bass liver lectin accounting for codon frequencies of striped bass. Degenerate primers FntermD (5'-CAA AGC TTT AYA ACT AYA ARA ACG TNG C-3'), RV3083D (5'-TCG AAT TCG TNA CGA TRT ANG GCT C-3'), RV437BD (5'-TCG AAT TCA CCT CAN CCT CRC A-3') and R3083D (5'-TCG AAT TCG TNA CGA TRT ANG GCT C -3') are used in RT-PCR amplification of liver cDNA. The amplification reaction consisted of 1 µl first strand cDNA, 1 µM each primer, 0.2 mM dNTPs, 1X Taq polymerase buffer, and 0.625 U of Taq polymerase (5U/µl) (Fisher Scientific) in a final volume of 25 µl held in 0.2 ml thin-walled tubes (Marsh Bio Products , Rochester, NY, USA). Thermal cycling was performed in a MJ Research DNA engine PTC-200 (Waltham, MA, USA). For MOPAC, the following parameters were used: (1) 72 °C for 20 s; (2) ramp 0.3 °C/s to 94 °C and hold 1 min; (3) 1.5 °C/s to 37 °C and hold for 1; (4) 0.2 °C/s to 72 °C and hold 1 min, repeat steps 2-4 five times; (5) 0.3 °C/s to 94 °C and hold for 1 min; (6) 0.7 °C/s to 50 °C and hold for 1 min+2s/cycle; (7) 0.7 °C/s to 72 °C and hold 1 min+4s/cycle, repeat steps 5 through 7 thirty nine times; fill-in at 72 °C for 10 min as described in [160]. As control for verifying the completion of the reverse transcription reaction, β-actin was amplified with ACTA.F (5'-TCA CCA ACT GGG ATG ACA TGG-3') and ACTB.R (5'-GAT GTC GAC GTC ACA CTT CAT-3'). Amplified fragments were extracted from the gel using a QIAquick gel extraction kit (Qiagen; Valencia, CA, USA) and TA-cloned into the pGEM-T easy vector (Promega; Madison, WI, USA).

Transformed bacteria were plated on Luria-Bertani agar prepared with 0.1 mM IPTG; 20 µg/ml X-gal; 100 µg/ml ampicillin and grown overnight at 37 °C.

Clones that contained the desired insert were identified by colony PCR [161]. Individual colonies were picked with automatic pipet tips and suspended in 50 µl of PCR grade-water.

The bacteria were lysed by heating in boiling water for 10 min. For amplification, 1 µl of the boiled bacteria was mixed in 20 µl of PCR reaction mix previously prepared [0.4 µM each primer, 0.2 mM dNTPs, 1X Taq polymerase buffer, and 0.5 U of Taq polymerase (5 U/µl)] (Fisher Scientific). PCR amplification was extended for 35 cycles.

E. Completion of full length cDNA

The hybrid bass lectin cDNA sequence was completed by rapid amplification of cDNA ends (RACE) using the Marathon cDNA amplification kit (Clontech Laboratories; Palo Alto, CA, USA). One µg of poly(A)⁺ RNA was reverse transcribed using the kit's anchored oligo(dT) primer. After second strand synthesis, an adaptor was ligated to the ends of the double stranded cDNA. Five µl of a 1/50 dilution of the adaptor-ligated cDNA was used as template for PCR amplification. The adaptor serves as an anchor site for the AP1 primer provided by the kit. The gene-specific primers used were 3083anti (5'-GAT GGA GGT GAC GAT GTA GGG CTC-3') for the 5' end and 3083sense (5'-AGC CCT ACA TCG TCA CCT CCA TC-3') for the 3' end. The reaction mix had the following composition 0.2 µM each primer, 0.2 mM dNTPs, 1X Taq polymerase buffer, and 1.25 U of Taq polymerase (Fisher Scientific) in a final volume of 50 µl. The cycling parameters were: (1) 94 °C for 1 min; (2) 94 °C , 30 sec; (3) 60 °C, 30 sec; (4) 72 °C , 2 min (repeat steps 2 through 4 30 times); fill-in at 72 °C for 10 min. PCR amplicons were cloned as described above.

F. DNA sequencing

Sequencing was performed in the Bioanalytical Services laboratory of the Center of Marine Biotechnology using an Applied Biosystems ABI 373 Stretch sequencer. Plasmids from clones chosen for sequencing were prepared by conventional alkaline lysis using the Wizard SV Plasmid Prep kit (Promega) following the manufacturers instructions. Each plasmid was sequenced from both ends of the cloning site using cycle sequencing with plasmid primers and Big-dye termination chemistry (Applied Biosystems). Contig assembly was accomplished using Sequencher v.3.1 software (Gene Codes; Ann Arbor, MI, USA).

G. Bioinformatics

Sequence manipulations were performed with Accelrys GCG Wisconsin Package v.8 (San Diego, CA, USA) operating on a Silicon Graphics IRIX workstation (Mountain View, CA, USA). Calculations of theoretical protein characteristics from the deduced peptide sequence were performed with ProtParam (www.expasy.ch). The signal sequence cleavage site was predicted [162] with the SignalP algorithm (www.cbs.dtu.dk/services/SignalP/).

Immune challenges

Fish (approximately 500 g) were injected over several postdorsal sites above the midline with 1 ml total volume of turpentine, and maintained in a 6 ft tank under the conditions described above. Control (uninjected) fish were only anaesthetized, and maintained under the same conditions as the injected fish. Fish were killed at 1, 3, 6, 12, 24, 48, 72, 96, and 120 h after injection, dissected, and the livers flash-frozen in liquid N₂.

A. Northern analysis

Ten µg of liver total RNA and 1 µg of poly(A)⁺ RNA, isolated as described above, were resolved by denaturing agarose formaldehyde gel electrophoresis [157, 161]. The RNA

samples were prepared by mixing 5 μ l of the sample with 15 μ l of Ambion Formaldehyde Load Dye (Austin, TX, USA) and incubated at 65 °C for 15 min and chilled on wet ice. To visualize the separation of the RNA in the gel, 1 μ l 200 μ g/ml ethidium bromide was added to the sample. The samples were loaded on a 1% agarose gel and separated at 3 V/cm electrode distance. Downward, passive capillary transfer following Ambion's bulletin 169 was used to transfer the resolved RNA to a membrane (www.ambion.com/techlib/tb/tb_169.html). The gel was placed on top of a charged nylon membrane, Zeta Probe GT (BioRad), which in turn rested on a stack of blotting paper (VWR Scientific; West Chester, PA, USA). A salt bridge was built with a strip of Whatman 3MM paper (Fisher Scientific) connecting a reservoir of transfer solution (3 M NaCl/10 mM NaOH) to the top of the transfer stack. The transfer was allowed to progress for 2 h after which the membrane was washed in 2X SSC (20X SSC: 300 mM sodium citrate (pH 7), 3 M NaCl). For storage, the membrane was air-dried and stored at room temperature sandwiched between blotting paper.

The synthesis of the radioactive probe used in membrane hybridization was performed by random hexamer-primed extension using template DNA prepared by PCR amplification from pFBP5 cDNA clone. The amplicon was gel-purified by agarose slab electrophoresis, and extracted with Qiagen's QIAquick kit. The double-stranded probe was produced in the presence of Amersham Redivue α [³²P]-dCTP with the Amersham RediPrime II kit, following the manufacturer's instructions. To remove the unincorporated nucleotide, the labeling reaction was loaded onto an Amersham ProbeQuant G-50 spin-column and processed as instructed by the manufacturer. To measure activity, 2 μ l of the purified probe was added to 3

ml of scintillation cocktail (National Diagnostics; Atlanta, GA, USA) and measured with the high energy channel of a Beckman LS 1801 scintillation counter.

Hybridization was performed in a forced-air oven/shaker (Amersham Biosciences) in rotating glass hybridization bottles. The dry transfer membrane was briefly wetted in 5X SSC and pre-hybridized in Ambion ULTRAhyb™ at 42 °C for 30 min, which contains 50% formamide. The DNA probe was denatured by heating at 100 °C for 10 min, followed by placing in wet ice. The probe was added to the pre-heated (42 °C) hybridization solution (70 µl/cm² membrane area), and mixed to achieve a probe concentration of at least 10⁶ cpm/ml. After hybridization overnight at 42 °C, the membrane was washed twice successively for 15 min at 65 °C in 2X SSC/0.1% SDS, 1X SSC/0.1%, and 0.5X SSC/0.1% SDS pre-heated at 65 °C, while monitoring the radioactivity remaining on the membrane with a handheld Ludlum Measurements Model 3 Geiger monitor (Sweetwater, TX, USA). The radioactive signal annealed to the membranes was imaged by autoradiography. First the washed membrane was blotted to remove excess liquid and wrapped in Fisherbrand® all-purpose laboratory wrap (Fisher Scientific). The membrane was sandwiched between a XAR film (Eastman Kodak; Rochester, NY, USA) and an intensifying screen, and exposure was carried out overnight at -80 °C. Development of the film was performed in an XOMAT automated developer (Eastman Kodak).

B. Western analysis

Serum samples (1/20) for SDS-PAGE were prepared with 2X reducing sample buffer.

Electrophoresis was performed with 10% acrylamide, 0.8% bis-acrylamide mini-format gels of 0.75 mm thickness. Resolved samples were transferred to PVDF (Millipore) in Towbin's buffer, 0.1% SDS as described above. Antiserum against liver lectin was produced in rabbits

by Duncroft (Lovettsville, VA, USA). Briefly, 100 µg lectin prepared in complete Freund's adjuvant was injected intramuscularly. On day 14 and day 28 a boost of 50 µg in incomplete Freund's adjuvant was injected subcutaneously. On day 56 the rabbits were bled, and antibody titers were tested by Western blot. A 1/6000 dilution of antiserum was prepared for Western blot in T-TBS/gelatin (0.1% Tween-20; 50 mM Tris-HCl, pH 7.5; 100 mM NaCl; 3% fish gelatin). Antibodies were detected with goat anti-rabbit antibodies conjugated to horseradish peroxidase (HRPO) (BioRad). Blots were developed with SuperSignal West Pico (Pierce Chemical), a luminescent HRP substrate, and imaged by autoradiography as described above.

Results and Discussion

Purification of a fucose-binding lectin

To screen for the presence of putative lectins in striped bass' liver a set of divinyl sulfone-conjugated carbohydrate affinity matrices were produced. Because lectins of various types may bind selectively to different sugars, Man, GlcNAc, Fuc, Glc and Lac were selected as potential ligands for the screening [163]. Most work [143, 164, 165] describing purification protocols for isolating hepatic lectins refer to earlier work on the purification of the asialoglycoprotein receptor from rabbit liver [142]. Their methodology describes using liver acetone extracts, which although a seemingly harsh treatment, facilitates protein purification from lipid-rich tissues (i.e. liver) by crudely precipitating the protein fraction. Thus, these methods were adapted for the purification of lectin(s) from fish livers. In the initial attempts, in spite of clarifying the sample by centrifugation, the chromatography columns clogged by a deposit left on the surface of the affinity matrix. These technical problems with slow or interrupted flow of the affinity columns led to attempts at batch affinity purification. However, no proteins could be eluted from the affinity matrices. Suspecting that acetone precipitation or the presence of detergents may have denatured the putative lectins, aqueous liver extracts prepared in the absence of detergent were tested for the purification procedures. Further, in an attempt to enrich the extract for lectin activity prior to the affinity purification, the crude homogenate was subjected to batch anion-exchange chromatography (AEX). Each of the fractions eluted (250 mM, 500 mM, 1 M NaCl) was loaded on the diverse affinity columns described above. From the Fuc-conjugated gel, a large peak absorbing at 280 nm appeared during elution.

From this initial purification a yield of 420 μg of lectin was isolated, which exhibited a single band of about 32 kDa (fig. 1) by SDS-PAGE in the presence of 2-ME. The absence of reducing agent resulted in a smaller band of 29 kDa suggesting that the lectin contained intra-chain disulfide bridges that upon reduction allow linearization of the polypeptide. The mobility of the bass lectin under reducing conditions corresponded to the size expected for a collectin subunit but the non-reduced size is unlike the characteristic trimer resulting from the covalent linking of their collagenous tail. To test if the bass lectin shared sequence similarity with collectins, the NH_2 -terminus was sequenced [YNYKNVAL(R)GKATQXA(R)YLX(T)(S)] and compared to the NCBI protein database, but no matches appeared.

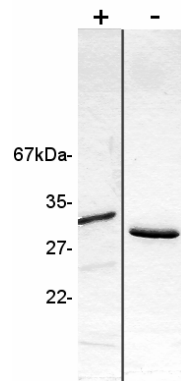


Figure 1. SDS-PAGE analysis of the hybrid bass hepatic lectin.

Lanes are presence (+) and absence (-) of 2-ME. Calibration markers are displayed on the left.

Attempts to reproduce the purification from liver tissues from other individuals failed to yield a comparable protein. As for many proteins produced by the liver, particularly as acute phase reactants during innate immune responses, it was assumed that the bass lectin was likely synthesized as a humoral component, destined for circulation in blood. Therefore, an attempt was made to isolate the lectin from serum collected from the hybrid bass. Using fresh batches of affinity matrix, diluted and dialyzed bass serum was directly processed by affinity chromatography. A protein eluting from the Fuc-conjugated column demonstrated the same

size as that isolated from the liver although other bands appeared also co-eluted. Thus, a second purification step using size exclusion chromatography was performed (fig. 2).

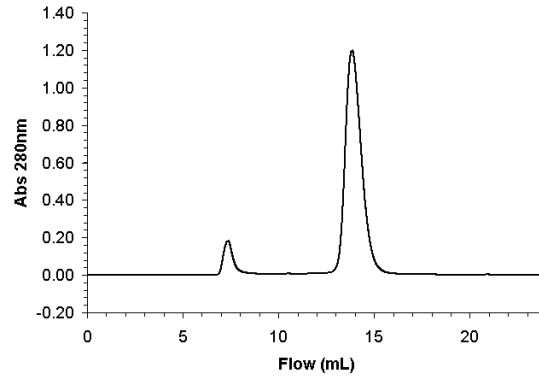


Figure 2. SEC polishing of Fuc affinity-purified serum lectin.

A larger retarded peak followed a small peak (10% total peak volume) eluting at the void volume of the column. This second chromatography step was enough to purify the lectin to homogeneity as judged by silver staining (fig. 3) and is referred hereon as Fuc-binding protein 32 kDa (FBP32).

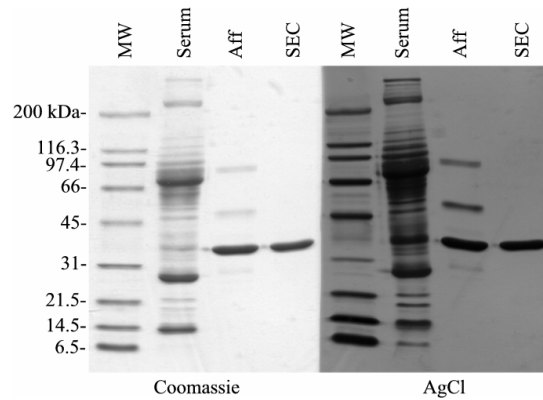


Figure 3. SDS-PAGE analysis of the purification steps for the bass serum lectin.

All samples were treated with 2-ME. Coomassie stain reflects relative band intensities while silver stain is for maximum sensitivity. Lanes are labeled according to the purification steps.

Several characteristics of the protein eluted in the void peak suggest that it is immunoglobulin μ (IgM), which is the main antibody isotype produced by fish. Dissociation of IgM in the presence of 2-ME produces a heavy chain (70 kDa) and light chains (25 kDa) [166]

corresponding to the bands co-eluted with the lectin. In addition, the exclusion limit of the column (500 kDa) is below the native size of IgM (>500 kDa). Interestingly, the presence of Fuc-specific IgM in serum suggests that the immune system recognizes this sugar as antigen. A faint band of approximately 50 kDa present in the affinity-purified fraction (fig. 3) was not identified in the SEC eluates. No further characterization of the IgM or of this minor band was performed; however, the detection of antibodies with identical specificity to the lectin in presumably unchallenged fish is indicative that the adaptive immune systems should be capable of responding to the same epitope. Future studies should address if these antibodies represent a natural subset, a phenomenon observed first in mammals but also present in ectotherms [167], or actually reflect a prior infection event that was effectively cleared.

SDS-PAGE analysis of the serum lectin produced similar results as for the hepatically isolated lectin (fig. 4). Although the bass serum lectin exhibits similar binding to the L-fuc affinity matrix and electrophoretic mobility as the hepatic lectin, their identification as the same protein required further confirmation by sequence from the NH₂-terminus of the serum lectin. The seven amino acids identified, Y-N-Y-K-N-V-A, matched those presented above from the

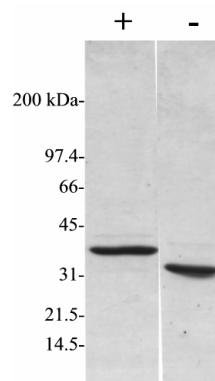


Figure 4. SDS-PAGE analysis of the serum form of the Fuc bass lectin. Separation was performed in the presence (+) and absence (-) of 2-ME. Calibration markers are displayed on the left.

hepatic lectin, providing more evidence to support the hypothesis that the serum lectin is synthesized by the liver and secreted to the blood stream. Accordingly, all of the following biochemical work was performed on the serum-purified lectin. Small-scale purification from striped bass produced a lectin identical to the palmetto bass, as judged by SDS-PAGE. Henceforth, the hybrid bass serum lectin will be referred to as MsxMcFBP32 and MsaFBP32 for the striped bass.

A more accurate mass estimation of the lectin was obtained by electron spray ionization-mass spectrometry (ESI-MS). The ESI-MS determined mass was 32.388 kDa (fig. 5) which matches closely the estimate calculated from SDS-PAGE.

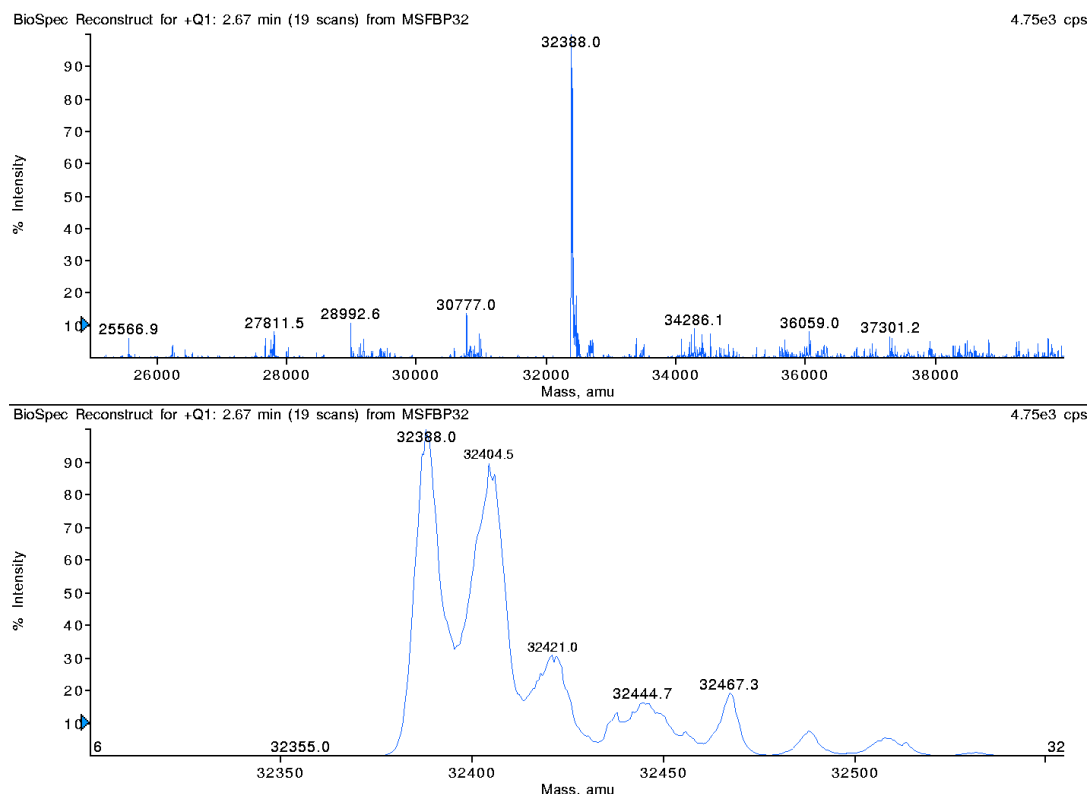


Figure 5. Determination of MsaFBP32 mass by ESI-MS. The lower panel is an expanded view of the most prominent peak (32388 kDa) in the upper panel.

Preliminary information on the hemagglutinating activity of the serum lectin was obtained by screening human red blood cells for potential ABO blood group specificity. All of the human ABO groups were agglutinated by the bass lectin with marked enhancement when treated with proteinase, which is in agreement with the presence of terminal Fuc α 1-2 on all types from this blood group system. Thus, human RBCs of blood group A were selected for preparing a purification table. After two purification steps, affinity and SEC, the bass serum lectin was purified 400-fold with a 97% recovery of total serum hemagglutinating activity (Table 1). From the purification table it is estimated that the lectin is present at $\sim 110 \mu\text{g}/\text{ml}$ in serum and makes up approximately 0.23% of total serum protein. Several subsequent purification experiments from hybrid and striped bass serum produced comparable results.

Table 1. Purification table of the bass Fuc-binding lectin from serum

Sample	Volume ^a	[Protein] ^b	Total Protein ^c	Titer ^d	Total Activity ^e	Specific Activity (S.A.) ^f	Yield ^g	Purification ^h
Serum	10	44.29	442.9	64,000	128,000,000	289,004	100	1
Affinity: Fuc	0.492	1.9	0.9348	128,000	12,595,200	13,473,684	9.84	47
SEC: Superose 12	2.04	0.295	0.6018	25,600	10,444,800	17,355,932	8.16	60
flow through	20	20.27	405.4	2,000	8,000,000	19,734	6.25	0.07

^aVolume of fraction (ml)

^bMeasured by microtiter Bradford (mg/ml)

^cVolume x protein concentration (mg)

^dDilution to +1/2 HA

^eVolume (ml)/5 ml *Titer (AU)

^fTotal activity/Total protein

^g% total activity

^hfold increase in S.A.

Characterization of the native FBP32

Results from SDS-PAGE and ESI-MS indicate that FBP32 consists of a single polypeptide but they do not provide information regarding the quaternary structure of the lectin in conditions closer to physiological state. To address this question, the size of the serum lectin in its native form was estimated by SEC using a calibrated column equilibrated in buffered saline. The elution position of the MsaFBP32 (fig. 6) under the chromatography conditions selected suggests an apparent molecular weight of 35.5 kDa, which is similar to the size determined under previous denaturing conditions.

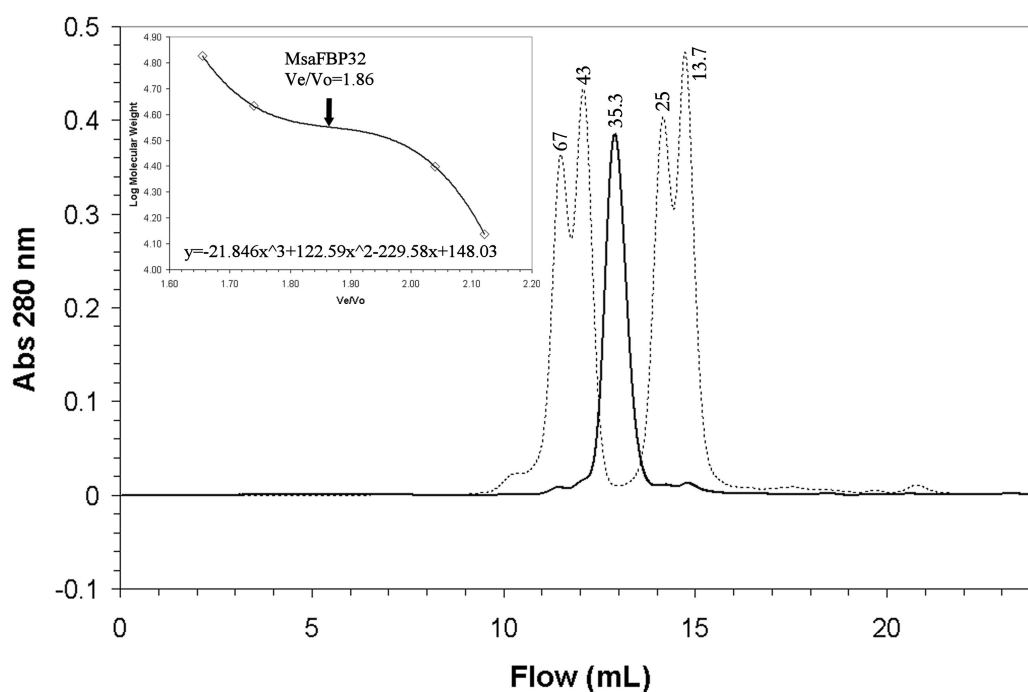


Figure 6. Size estimation of MsaFBP32 by SEC.

Overlay of independent chromatographic traces for calibration markers (dashed line) and for MsaFBP32 (solid line). Numbers indicate molecular weight in kilodaltons. Inset. Calibration curve of markers indicating position of elution of lectin.

Similar results were obtained from cross-linking the purified lectin with a water-soluble, homobifunctional cross-linker that is not cleaved by 2-ME (fig. 7). Stoichiometrically

increasing amounts of BS3 produce an ever increasing number of larger molecular species which are separated by intervals corresponding to the monomers weight. This incremental pattern most likely reflects artifactual polymerization of the lectin rather than a natural association between subunits.

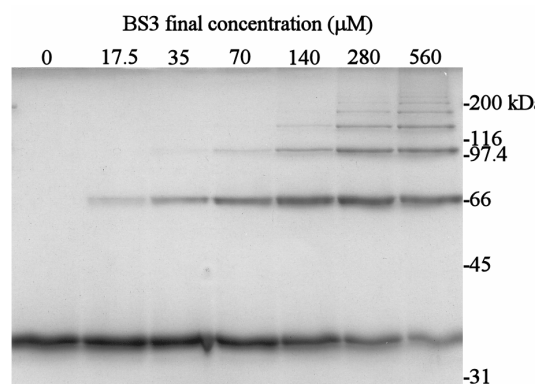


Figure 7. SDS-PAGE of MsaFBP32 cross-linked with BS3.

The final concentration of cross-linking reagent in each reaction is indicated at the top of each lane. The gel used was 10% T and was stained with Coomassie blue R-250. Migration of molecular weight calibration markers is indicated at left.

The consensus from these independent size determinations is that FBP32 naturally occurs as a 32 kDa monomer possessing internal disulfide bridges that likely contribute to its structural stability in circulation. This contrasts with the other reported fish serum lectins, which either associate into dimeric [126] or multimeric molecules [127, 168, 169].

Isoelectric focusing analysis

Although electrophoresis and chromatography indicate that FBP32 is pure, the heterodispersity of peaks observed in ESI-MS suggest that isoforms may be present. To investigate this possibility and further characterize the lectin's biochemical features, isoelectric focusing (IEF) was performed in the presence of urea as a denaturant. The IEF-resolved MsaFBP32 revealed limited heterodispersity (fig. 8) with one strongly stained band (pI 4.97) and four weaker bands (pI 5.09, 4.91, 4.87, and 4.86) all exhibiting the same electrophoretic

mobility when analyzed by 2D-PAGE (fig. 9). Analysis on native IEF revealed higher pI than the denatured protein (data not shown). An assay for detection of glycosylation did not detect any carbohydrate modifications on MsaFBP32 (data not shown) excluding any possibility that microheterogeneity of glycans is the cause of the observed heterodispersity. Presence of isoelectric point heterodispersity has been reported for other lectins [146] and it is plausible that these represent sequence isoforms or post-translational modifications. The reproducibility of the band pattern has not been tested, but is possible that these minor bands actually represent artifacts due to the deamidation of basic residues or carbamylation from the urea present during electrophoresis.

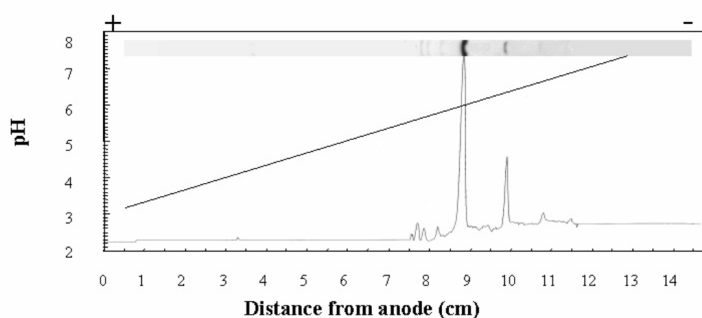


Figure 8. Denaturing isoelectric focusing of MsaFBP32

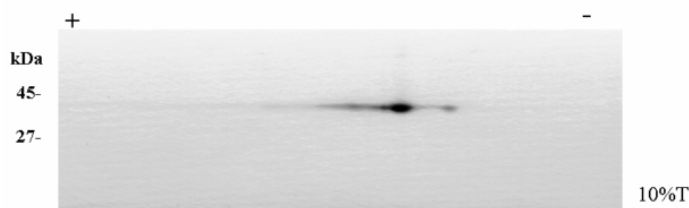


Figure 9. 2-D PAGE analysis of SEC-grade MsaFBP32.

Carbohydrate specificity

From the screening with affinity columns, it became evident that MsaFBP32 recognizes Fuc (i.e. 6-deoxy-L-galactose) preferably over Man, Glc, GlcNAc or Lac. However, in circulation

Fuc is mostly protein-bound [170] suggesting that the lectin unlikely encounters significant amounts of monosaccharide. To determine the configurational constraints required for lectin binding to carbohydrate ligands, diverse carbohydrates were tested for their inhibitory activity on agglutination (fig. 10). Among these were included glycans and glycoconjugates possessing Fuc that may represent biologically relevant ligands. As expected, Fuc, fucosyl-oligosaccharides and glycoproteins rich in Fuc are the preferred ligands for MsaFBP32 (table 2). In contrast to most other natural hexopyranoses, fucose is predominantly in the L-configuration and is a 6-deoxy sugar [171]. The relatively lower inhibitory capacity of D-Fuc (4C_1) and the even weaker effects of other D-hexopyranose isomers illustrate the importance of the 1C_4 conformation adopted by Fuc to binding MsaFBP32. Comparison of methyl glycosides of Fuc indicates a preference for the α anomeric configuration over β . The difference is further enhanced when a large bulky substituent (i.e. *p*-nitrophenyl) is the aglycone. Because L-sugars are enantiomers to D-sugars the anomeric configurations are reversed, which means α is above the plane of the Fuc ring when viewed in standard orientation. The importance of the absence of hydroxyl group on C6 in Fuc is demonstrated by the relatively weaker inhibitory activity of L-Gal. Further, the increased hydrophobicity of deoxy C6 apparently contributes to the binding of Fuc. Competition with fucosylamine, which has an amino modification on C2, showed comparable levels of inhibition to L-Gal suggesting that loss of C2 hydroxyl is important but does not abrogate binding. The effect of modifications on C3 and C4 of Fuc on binding was not tested but weak inhibition observed by Man may provide a clue to their role in specificity. Structural studies in C-type lectins described [172] how the hydroxyls on C2 and C3 of Man, which adopt the same configuration as hydroxyls on C4 and C3 of Fuc, are able to bind albeit in a different orientation. It is likely

that weak inhibition of MsaFBP32 by Man reflects a similar phenomenon supporting the importance of a (+) *syn*-clinal configuration in the C3-C4 bond as present in Fuc.

In oligosaccharide chains of glycoconjugates, such as human blood group substances, Fuc is commonly found as the terminal residue linked to C2, C3 or C6 of the subterminal sugar. To test if MsaFBP32 bound Fuc in the different linkages present in these natural glycans, inhibition was performed with oligosaccharides present in ABO group (H-antigen), Lewis group (2'-fucosyllactose and 3-fucosyllactose) and the N-glycan core modification, fucosyl(1-6)N-acetylglucosamine. Similarly, glycoproteins described to contain fucosylated glycans such as gastric mucins [173, 174] were tested as inhibitors. All four fucosylated oligosaccharides inhibited agglutination, indicating that its terminal position is sufficient for binding. The presence of a subterminal GlcNAc appeared to enhance inhibition in spite of this monosaccharide not exhibiting significant inhibition by itself. As expected, the fucosylated glycoproteins were effective inhibitors despite removal of terminal sialic acids. Interestingly, no inhibition of agglutination was observed with algal fucoidan in spite of being a fucose polymer. Chemical analysis of fucoidan reveals it consists of 1-3 linked Fuc core substituted at C4 by sulphate or branching fucose residues on some of the residues [175]. Hence, the presence of unsubstituted hydroxyls at C3 and C4 of Fuc appear necessary for the lectin to bind. In summary, the 1C_4 chair conformation placing free hydroxyls axial at C-4 and equatorial at C-3, the axial C1 in the α anomer, the deoxy C-6 of Fuc, all appear to contribute to the preference for this ligand.

The affinity of MsaFBP32 for Fuc is shared with lectins described in bacteria, plants, and animals. However, structures resolved from bacterial lectins [176, 177], a plant lectin [178],

and a fungal lectin [179, 180] clearly indicate that binding to Fuc can be achieved through topologically unrelated folds. Plant and animal lectins appear to have emerged independently during evolution. Thus, considering the source of MsaFBP32 its ancestry probably lays with the latter. Two families of animal lectins that bind Fuc have been described: C-type lectins and homologous lectins from the Japanese horseshoe crab [181], and a fuclectin from eel serum [182, 183]. Nevertheless, in C-type lectins that bind Fuc, Man and GlcNAc are also bound equally well since all three monosaccharides share a (+)*syn*-clinal hydroxyl configuration. In contrast, the specificity of the eel and horseshoe crab lectins is restricted only to Fuc. Several other Fuc-specific lectins from fish serum have been reported [184], but their primary sequence was not determined so their ancestry remains unknown. Structural models of C-type lectins [39] explain the interchangeability of ligands in the binding site but no such detail is available to explain the specificity expressed by the eel and horseshoe crab lectins.

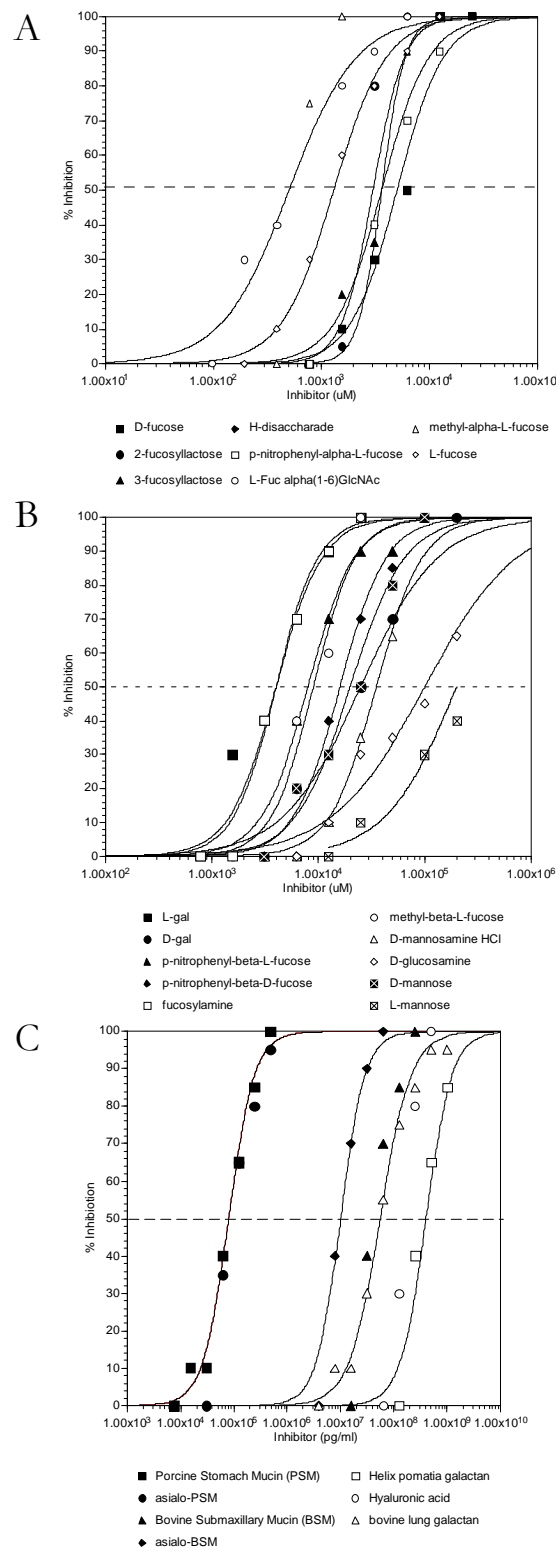


Figure 10. Hemagglutination inhibition curves for A. monosaccharides, B. polysaccharides and C. glycoproteins

Table 2. Hemagglutination inhibition profile of hybrid bass FBP32 by monosaccharides, oligosaccharides, and glycoproteins*

Inhibitor	50 % Inhibition Concentration ¹	Specificity Factor ²
	(mM)	
Fuc α (1-6) GlcNAc	0.5	2.60
Methyl α -L-Fuc	0.7	1.86
L-Fuc	1.3	1.00
2'-fucosyllactose	2.5	0.52
3-fucosyllactose	3	0.43
H-dissacharide	3.6	0.36
<i>p</i> -nitrophenyl α -L-Fuc	3.6	0.36
L-Gal	4	0.33
fucosylamine	4	0.33
D-Fuc	5	0.26
<i>p</i> -nitrophenyl β -L-Fuc	8	0.16
Methyl β -L-Fuc	9	0.14
<i>p</i> -nitrophenyl- β -D-Fuc	16	0.08
D-Man	20	0.07
D-Gal	25	0.05
mannosamine HCl	35	0.04
glucosamine	100	0.01
	(μ g/ml)	
Porcine gastric mucin (PGM)	0.08	1
asialo-PGM	0.08	1
Bovine submaxillary mucin (BSM)	40	0.0020
asialo-BSM	10	0.0080
Bovine lung galactan	55	0.0015
Hyaluronic acid	160	0.0005
Helix pomatia galactan	400	0.0002

*Mono- and oligosaccharides not inhibitory at 200 mM: D-Glc, D-glucosamine, GlcNAc, N-acetyl glutamine, N-acetyl galactosamine, N-acetyl mannosamine, N-acetyl muramic acid, N-acetyl glutamic acid, and L-Man. Glycoproteins and polysaccharides not inhibitory at 1 mg/ml: fetuin, asialo-fetuin, heparin and chondroitin sulfate and fucoidan.

¹Concentration of carbohydrates and glycoproteins required to inhibit 50% of the agglutinating activity (IC₅₀) as scored in assay. Pronase-treated human A RBC were used as test cells.

²Calculated in relation to Fuc and PSM, respectively.

Cation requirements

The role of the presence of metals in lectins ranges from principally structural [185] to mediating directly interaction between protein and the saccharide ligand [39]. Calcium directly binds to the saccharide ring hydroxyls in C-type lectins [39] and *P. aeruginosa* lectin [176, 177], and therefore it is essential for maintaining their activity. However, in no other animal lectin families does calcium directly interact with simple monosaccharides, although lectin activity is dependent on the presence of calcium in pentraxins [40] and fuclectins [181, 182]. Addition of 30 mM EDTA to MsaFBP32 bound to a Fuc affinity column did not elute the lectin suggesting its activity was not dependent on cations (data not shown). More extensive treatment of the lectin with EDTA under acidic conditions failed to affect its hemagglutination activity. Neither did the addition of increasing concentrations of Ca^{2+} to the assay improve agglutination titers. Thus, the lectin activity of bass FBP32 does not appear to require divalent cations. The lack of strong affinity for the typical ligands (i.e. GlcNAc and Man) of the Fuc-binding C-type lectins and the apparent lack a cation requirement for binding strongly suggest that the bass Fuc-binding lectin does not belong to this family. However, to identify with certainty if the bass lectin belongs to any of the characterized lectin families or is unique will require deducing its peptide sequence.

Cloning of the hybrid bass lectin

To determine the nucleotide sequence of the bass lectin transcript, the MOPAC procedure [159] was employed to obtain an initial PCR amplicon, which could be later extended to complete the transcript. The initial step was to obtain peptide sequence of the NH_2 -terminus and several internal peptides, on which the degenerate primers needed for amplification could

be designed (Table 3). Because the relative positions of the internal peptides in the polypeptide sequence were unknown, a forward primer was designed from the N-terminal peptide sequence and reverse primers on the internal peptides, all selected with the lowest codon degeneracy possible.

Table 3. Edman sequencing of bass hepatic lectin

Peptide name	Digestion enzyme	Peptide sequence
NH ₂ -terminal	-	YNYKNVAL(R)GKATQXA(R)YLX(T)(S)
3081	Trypsin	NSDFEAGSCTHTIEQTNPX(S)
3082	Trypsin	YVTVLLPGTNK
3083	Trypsin	VDLLEPYITSITIT(N)(R)
V437	Trypsin	YVNIVIPGREEYLTLCVEVEYGSVL(L)
V438	Trypsin	ATQSSLFESGIAYNAIDGNQAN(NNWEMASETH)
	Lys-C	(K)NTMNP

Residues in parenthesis are tentatively assigned

After using several combinations of primers, the longest PCR amplicon (861 bp) was obtained with primers FNTERMD and RV437BD (fig. 11). Upon sequencing this amplicon, the deduced peptide sequence was found to contain all of the other sequenced peptides (i.e. 3081, 3082, 3083, v438 and Lys-C).

Although the sequence obtained by MOPAC accounted for all of the sequenced peptides, the full open reading frame (ORF) was not complete. The complete ORF was obtained by RACE in both directions. With forward primer 3083sense.F, a 917 bp amplicon (fig. 12, lane 2) was obtained for the 3' end RACE reaction. For the 5' end the cDNA, the 3083anti.R reverse primer produced a 340 bp amplicon (fig. 12, lane 3). The assembly of all three amplicons resulted in a contig of 1,099 bp (fig. 13). The ORF deduced from this contig was judged to be complete since a stop codon was present upstream of the putative start methionine codon. Several stop codons were also present downstream of the initial stop.

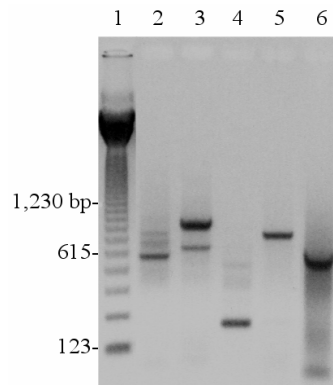


Figure 11. PCR amplification of MsxMcFBP32 transcript by MOPAC. Lanes: 1) 123 bp marker, 2) FNTerm/Rv438DB primers, 3) FNTerm/Rv437DB primers, 4) FNTerm/R3083D primers, 5) FNTerm/R3082D primers, 6) ACTA.F/ACTB.R positive control primers. Gel: 1.5%, 1X TAE

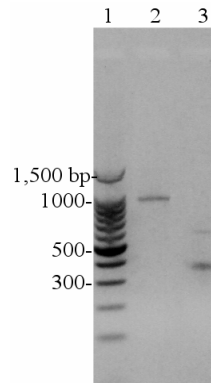


Figure 12. 5' and 3' RACE of MsxMcFBP32 cDNA. Lanes: 1) 100 bp marker, 2) 3083sense/AP1, 3) AP1/3083anti. Gel: 1.5%, 1X TAE

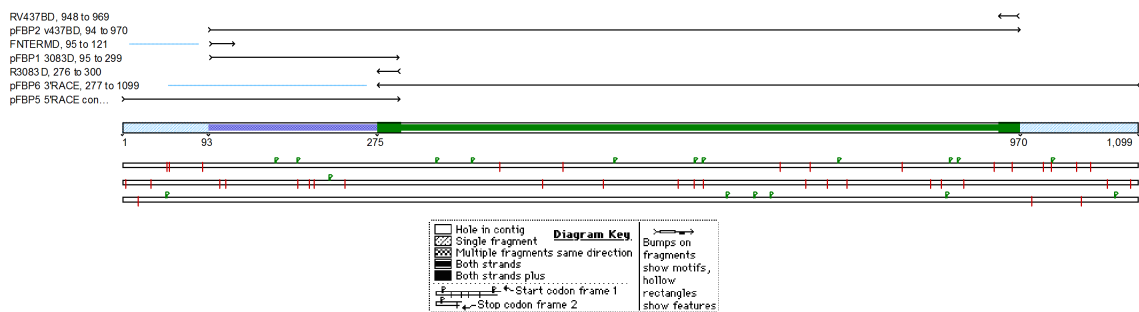


Figure 13. Sequencing contig scheme of the full-length MsxMcFBP32 cDNA. Clones pFBP2 and pFBP1 represent the amplicons obtained by MOPAC using degenerate primers based on the peptides sequenced by Edman's degradation (short arrows). The clones pFBP6 and pFBP5 are the 3' and 5' RACE amplicons, respectively. Rectangles below represent the three forward reading frames of the contig. Within the reading frames, vertical lines represent stop codons and "P" represents a methionine start codon.

The ORF encoded a 311 residue polypeptide, of which the first 18 amino acid residues corresponded to a signal peptide accurately predicted by the SignalP algorithm [162] as confirmed by the NH₂-terminal peptide sequence (fig. 14). In retrospect, 37% of the bass hepatic lectin had been completed by Edman sequencing. The calculated mass (ProtParam) for the 293 residue mature polypeptide was 32,449.2 Da, in good agreement with the biochemical estimates. The 61 Da mass discrepancies between the calculated mass and the mass determined by spectrometry (32,388 Da) may reflect unknown polymorphism(s) or secondary modifications. Flanking the ORF, the cDNA contained a 47 bp 5' untranslated region (UTR) and a 110 bp 3' UTR ending in a canonical polyadenylation signal. Interestingly, a pairwise comparison (fig 15, Panel A)) of the deduced polypeptide revealed internal sequence duplication. This is not unusual among lectins since other tandemly arranged domains have been described in unrelated lectin families such as: macrophage mannose receptor [186], galectin-4 [187] and galectin-8 [188], and in less described groups: tachylectin-1 and 2 [189, 190], a tunicate lectin [191] and the steelhead trout egg lectin [192]. Notably, the deduced polypeptide of FBP32 did not show any similarity to the CRDs of any known lectin families to date [31-33, 193] or to the collagen tails characteristic of collectins. However, a gene encoding an MBL-like protein was previously cloned from the zebrafish and common carp [125], demonstrating that teleosts do possess collectins. Unlike the mammalian MBPs, however, based on its CRD motif the cyprinid homologue is predicted to bind Gal rather than Man and is expressed in spleen and not liver. It is of note then, assuming this protein is broadly distributed among fish taxa, that during the affinity screen of bass humoral lectins no lactose binding-protein matching the size of the cyprinid MBL was encountered. Therefore, it is unclear if this protein is present in the serum of the striped bass or of cyprinids.

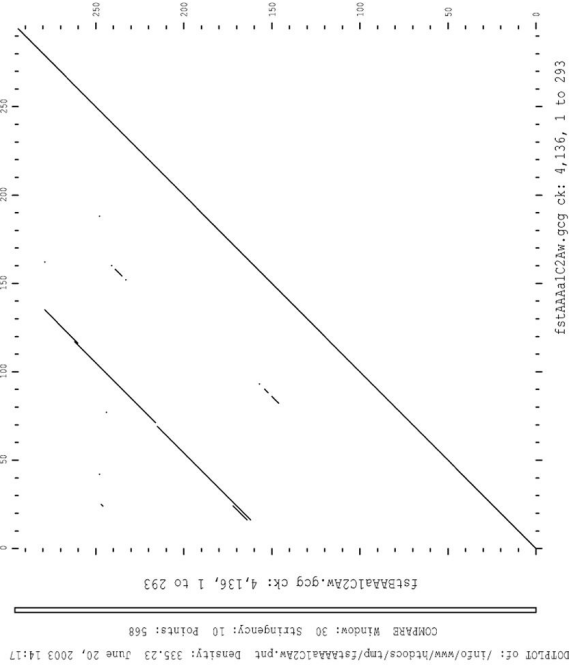
Each duplicated motifs is approximately 140 residues long (fig. 15, Panel B). The duplicated domains of bass lectin ranged from Asn₅ to Gly₁₄₅ (N-domain) and from Asn₁₅₃ to Gly₂₈₉ (COOH-domain) covering most of the mature protein. A 7 residue spacer links the two putative domains. Pairwise comparison using the Blosom 62 matrix indicates the domains share 50% identity and 67% similarity. The considerable length of these motifs suggest they encode a structural fold or domain which mediates carbohydrate binding. The distribution of cysteine residues within each motif is noteworthy (fig. 15, Panel B), especially the unusual contiguous cyteines in the N-domain. Due to the closeness of these residues in the polypeptide chain they most likely form a disulfide bridge between them, though it remains to be demonstrated biochemically. The presence of solely intrachain disulfide bridges, as deduced from SDS-PAGE, is supported by the presence of an even number of cysteine residues within each motif. In lepidopterans [68, 194] and an ascidian [191], a humoral lectin with distinct tandem CTLDs has been identified which are implicated in immunity. No tandem CTLDs have been identified in vertebrates, so the distinct lectin domains of FBP32 may perform in vertebrates a function analogous those of the invertebrate tandem CTLDs.

1 CTGGACTCCAGGGATAAAAGATCTGTTCTAACCAGGAAGCAGGAATAATGAGGCACAGTGTGGTATTTCT
m r h s v v f l -11
71 GTTGCTGCTCCTCTTAGGGGCGTGTTCACTTACAACATAAAAAATGTGGCCTTGCGTGAAAAGCGACT
l l l l l g a c s a Y N Y K N V A L R G K A T 13
NH₂-terminal
141 CAGTCGGCACGTTATTTGCACACACATGGAGCCGCTACAACGCCATTGATGGAAACCGTAACCTCTGACT
Q S A R Y L H T H G A A Y N A I D G N R N S D 36
3081
211 TCGAAGCTGGATCGTGACCCACACTATTGAACAGACCAACCCCTGGTGGAGAGTGGACCTACTGGAGCC
F E A G S C T H T I E Q T N P W W R V D L L E P 60
3083
281 CTACATCGTCACCTCCATCACCATACCAACAGAGGAGACTGCTGTCCAGAAAGGCTCAACGGGGTGGAG
Y I V T S I T I T N R G D C C P E R L N G V E 83
351 ATTCACATCGGCAACTCTATACAAGAAAATGGTGTGCAAACCCAAGGGTGGTGTAAATTTCTCATATCC
I H I G N S I Q E N G V A N P R V G V I S H I 106
421 CTGCAGGGATCTCACATACTATCAGTTTCACTGAACGTGTGGAGGGACGTTACGTGACTGTGCTTCTACC
P A G I S H T I S F T E R V E G R Y V T V L L P 130
3082
491 TGGTACAAACAAGGTTCTTACACTCTGTGAAGTGGAGGTTTCATGGGTACCGAGCCCCAACTGGAGAGAAC
G T N K V L T L C E V E V H G Y R A P T G E N 153
561 CTGGCCCTCCGAGGAAAAGCCACACAGTCTTCATTGTTTGAATCTGGTATTGCATATAATGCCATTGATG
L A L R G K A T Q S S L F E S G I A Y N A I D 176
V438
631 GGAATCAAGCCAACAATTGGGAAATGGCCTCCTGCACTCACACAAAAACACAATGAACCCCTGGTGGCG
G N Q A N N W E M A S C T H T K N T M N P W W R 200
Lys-C
701 AATGGATCTGAGCAAAACCCACAGAGTGTCTTCTGTTAAGGTAACCAACCGAGATTCATTTGAAAAACGA
M D L S K T H R V F S V K V T N R D S F E K R 223
771 ATCAATGGAGCTGAGATCCGAATTGGAGATTCCCTCGACAACAACGGCAACAACAATCCCAGGTGTGCTG
I N G A E I R I G D S L D N N G N N N P R C A 246
841 TGATCACAAGCATCCCAGCAGGTGCTTCTACTGAATTCAGTGTAACGGGATGGATGGCCGCTATGTAA
V I T S I P A G A S T E F Q C N G M D G R Y V N 270
V437
911 CATTGTTATCCCTGGAAGAGAAGAGTACCTGACCCTGTGTGAGGTGGAGGTGTATGGCTCTGTCTGGAT
I V I P G R E E Y L T L C E V E V Y G S V L D 293
981 TAGGTGTCACTACTAATACTGTTGAATGTACACAAAACAAAACAAAATAGTAGATTAAGCTTTTTTGATTG
*
1051 TTTCCATTCAAAATAAGACAGAGATGGTCTTATCCAATAAAATTACATCACG

Figure 14. The complete cDNA and deduced protein sequence of MsxMcFBP32.

The cleaved signal peptide is indicated by lowercase amino acids and is negatively enumerated. Chemically sequenced peptides are underlined with a single line. The in-frame stop codon is marked with an asterisk. The polyadenylation motif is double underlined. The duplicated domain is highlighted in yellow.

A



B

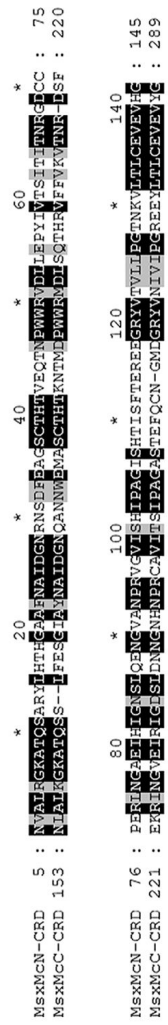


Figure 15. Duplicated motif present in the deduced sequence of MsxMcFBP32.

A. Dot plot (GCG Wisconsin Package) presenting offset diagonal (upper left corner) indicating similarity between NH₂-end and COOH-end of the polypeptide. Scale on abscissa and ordinate axis is position in peptide. **B.** Pairwise alignment of the duplicated peptide motif. Residue numbering is at both ends of each motif. Identity is shaded in black. Similarities (Blosum 62 matrix) are shaded in gray.

Expression analysis of FBP32

Demonstrated by its qualitative and quantitatively reproducible purification from blood, FBP32 appears to be a constitutively expressed protein leading to relatively invariant levels in circulation (fig. 16). However, the difficulties encountered while purifying FBP32 from the liver raised the possibility that, in analogy of mammalian APRs, the presence of lectin was correlated with inflammation brought on by either stress or disease. The FBP32 cDNA gene sequence made available the probe necessary for testing the levels of lectin transcripts in liver after experimental inflammatory challenge induced by intramuscular injection of turpentine, a reagent commonly for this purpose [195]. After hybridization to liver RNA obtained at different times post-injection, the hybridization signal was normalized with the intensity of the constitutive 28S rRNA band stained with ethidium bromide (fig. 17). Analysis of normalized band intensity indicates no difference in lectin transcript levels among treated fish and controls prior to 48 h post-injection, although levels increased approximately 2-fold at 72 h post-injection in treated fish. Similarly slow kinetics of expression of an APR (i.e. serum amyloid A) after injection with bacterial pathogen has been documented in salmonids [122]. However, the control fish demonstrated that FBP32 is normally expressed in the absence of inflammatory challenge. Similar results were observed for plasma levels of the Japanese eel fucoselectin [182]. Only after challenging primary hepatocyte cultures with lipopolysaccharide did levels of transcript and protein increase. Despite the presence of homologues of the mammalian acute phase reactants, such as pentraxins, it is possible that in poikilotherms the kinetics of systemic inflammatory responses is not as rapid as in mammals. This is further supported by results demonstrating that CRP levels in *X. laevis* do not rise as expected after turpentine challenge [196]. In addition, development of an antibody response in teleosts is much slower and

extended relative to that in homeothermic animals, which may reflect a temperature dependence of these processes [112]. An important problem in studying the APR using non-human models is that there is great interspecies variability in the changes of each APP [22]. Moreover, the response even varies between closely related rodents such as rats and mice. In other words, despite individual APRs being widely conserved, regulation of their expression is unique in each mammalian lineage and for that matter in vertebrates in general. Therefore, it should not be surprising if the profile of APRs expressed in ray-finned fish, if such a response exists, differs even among the highly diverse teleostean families. This initial assessment of FBP32 expression is only preliminary, but it appears that it does not respond like the typical mammalian acute phase reactants.

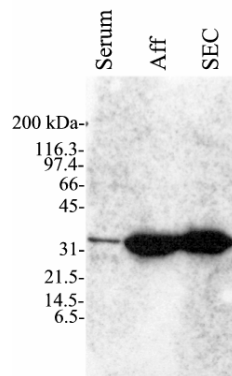
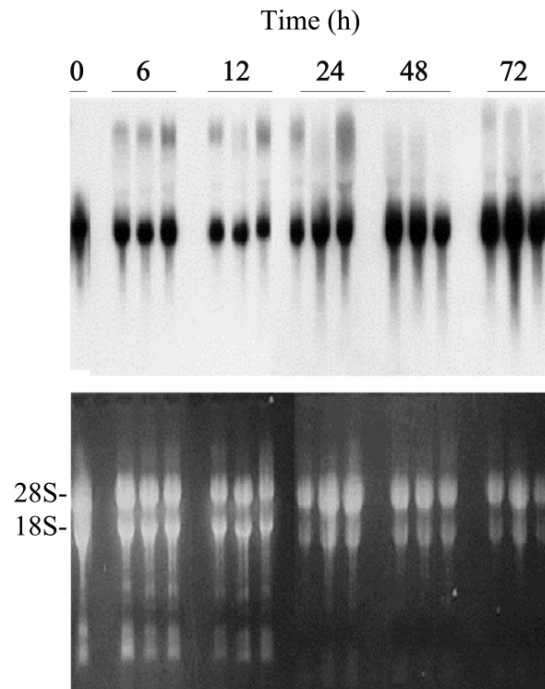


Figure 16. Immunoblot detection of MsaFBP32 during purification.

In summary, a 32-kDa lectin from bass liver and blood showing specificity for Fuc was purified to homogeneity. The deduced polypeptide sequence of FBP32 indicates it is not a member of the C-type family or any other lectin family described so far, and therefore represents a novel lectin within vertebrates. The duplicated 140-peptide motif found in FBP32 likely represents individual folding domains, which bind sugar, allowing the lectin to crosslink glycoconjugates. Expression of FBP32 in the liver as a systemic response to experimental

inflammatory challenge appears to be delayed relative to the local inflammatory response, although dynamics of homeostatic processes are not fully understood in cold-blooded animals.

A



B

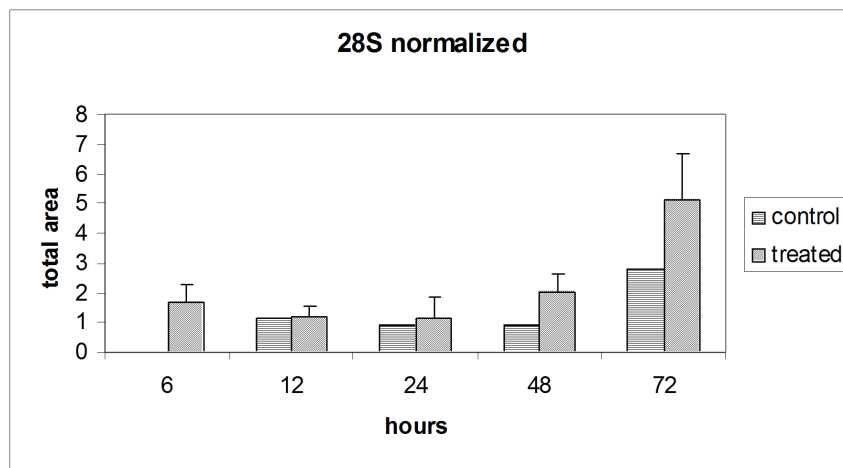


Figure 17. Northern analysis of FBP32 in liver after turpentine challenge.

A. Hybridized blot (above) and EtBr-stained electrophoresis gel (below). **B.** Normalization of hybrid blot using ribosomal band intensity.

CHAPTER 3. EMERGENCE OF A NOVEL LECTIN FAMILY

The search for an MBP/selectin-like lectin from fish led to the isolation and characterization of a unique Fuc-binding lectin with characteristic tandem domains. The lack of any similarity to motifs described for other lectins suggests that it may represent a novel animal lectin family still undescribed. During the 1990's, advances in technology led to an acceleration in the accumulation of DNA sequences deposited in gene databases. The potential of genomics in explaining organismal complexity has led to an ever-expanding list of organisms being represented in the databases rather than just the conventional model organisms. Meanwhile, refinement of computational tools for sequence comparison and improved accessibility allow exploiting this accumulating resource of DNA sequence. The combination of these has not only led to an expansion in recognizable members of known protein families but also in identifying new families. This in turn has allowed identification and assessment of the functional relevance of conserved sequence motifs or particular residues. Accordingly, determining ancestry of the bass lectin may facilitate the identification of residues that are functionally relevant. This chapter discusses the identification of bass lectin homologues present in disparate taxa, which illustrate a lectin family of great pliability yet surprising paucity in tetrapods.

Materials and Methods

Cloning FBP32 cDNA from parental species

To clone the full-length cDNA of FBP32 from the hybrid bass parental species, RT-PCR was performed using liver cDNA from *M. saxatilis* and *M. chrysops*. Liver total RNA was prepared,

as indicated in Chapter 2, and converted to cDNA by reverse transcription with the following modification from the procedure used in Chapter 2. MCSRACE_{T₁₆}VN, a lock-docking cDNA-synthesis primer (5'-CCG CAT GCG GCC GCA GAT CTA GAT ATC GAT TTT TTT TTT TTT VN-3') [197], was substituted for the previously mentioned oligo(dT) primer. This protocol modification allows using the MCSRACE primer (5'-CCG CAG ATC TAG ATA TCG A-3') for 3' RACE. The primers used for RT-PCR were FBP1.F (5'-CTG GAC TCC AGG GAT AAA AGA TCT G-3') and MCSRACE. The primer FBP1.F was designed to anneal to the beginning of the MsxMc full-length cDNA, and therefore should amplify the full-length parental species cDNA when used in combination with MCSRACE. To ensure fidelity of the amplicons, Invitrogen's eLONGase™ (San Diego, CA, USA), a proofreading polymerase mixture, was used for PCR. The PCR reaction mix consisted of 1X eLONGase™ buffer A, 1X eLONGase buffer B, 0.2 mM dNTPs, 0.2 μM of each primer, 1 μl of cDNA, and 1 μl of eLONGase polymerase mix in a 50 μl reaction volume. The cycling parameters followed the touchdown PCR technique [198], where the annealing temperature is lowered successively at each cycle so as to increase specificity of the reaction. Cycling parameters are: (1) 94 °C for 1 min; (2) 94 °C, 30 sec; (3) 60 °C, 30 sec minus 0.5 °C per cycle (4) 68 °C, 2 min (repeat steps 2 through 4 for 19 cycles); (5) 94 °C, 30 sec; (6) 50 °C, 30 sec; (7) 68 °C, 2 min (repeat steps 5 through 7 for 19 cycles); (8) fill-in at 68 °C for 3 min. Amplicons were ligated into a plasmid vector, and sequenced as described in Chapter 2.

Gene database search for homologues

To identify proteins with similarity to the FBPL motif described for MsxMcFBP32, a search of publicly available gene repository databases was undertaken. The query sequence used for

searching was the deduced peptide sequence of the N-domain of MsxMcFBP32 from Asn₅ to Val₁₄₅. Searches of protein database were performed with the BLASTP algorithm and of DNA databases with TBLASTN algorithm [199] residing on the web sites' server (www.ncbi.nlm.nih.gov/BLAST/). A search of dbEST [200] was performed with the TBLASTN algorithm. Searches of the pufferfish genome of *Fugu rubripes* were performed at the Department of Energy's Joint Genome Initiative (www.jgi.doe.gov) and for *Tetraodon nigroviridis* at the Genoscope (www.genoscope.cns.fr/externe/tetraodon/). A search for the lectin domain in Urochordates was performed using the whole genome shotgun-sequencing traces stored at *Ciona intestinalis* web page (ghost.zool.kyoto-u.ac.jp/indexr1.html).

Cloning and sequencing of Xenopus laevis FBPL ESTs

Confirmation of the identity of the *X. laevis* liver EST (GenBank Accession BE508834) matching PXN1-XENLA was accomplished by 3' RACE. The primer XL1.F (5'- AAG GTC TTC TCA ATT GCG GTG A-3') was used in amplifying the transcript from *X. laevis* liver total RNA that was reverse transcribed with the lock-docking primer as described above. The PCR reaction mixture was: 1X Takara buffer, 2 mM dNTPs, 0.4 mM each primer, 1.25 U Takara rTaq polymerase (PanVera Corp., Madison, WI, USA) for a 50 µl final reaction volume. The touch-down PCR cycling parameters were: (1) 94 °C for 3 min; (2) 94 °C , 30 sec; (3) 65 °C, 30 sec minus 0.5 °C per cycle (4) 72 °C , 3 min (repeat steps 2 through 4 for 19 cycles); (5) 94 °C, 15 sec; (6) 55 °C, 15 sec; (7) 72 °C, 3 min (repeat steps 5 through 7 for 24 cycles); (8) fill-in at 72 °C for 5 min.

For 5' RACE, Invitrogen's GeneRacer kit (San Diego, CA, USA) was used according to the manufacturer's instruction. First, cDNA produced from liver total RNA was PCR amplified

using primer XL3.R (5'-TGC TCA AAA TCA CCA TTC CTC TC-3') and the kit's 5' adapter primer. The PCR reaction mixture was: 1X Takara buffer, 2 mM dNTPs, 0.2 mM each primer, 1.25 U Takara rTaq polymerase (PanVera Corp., Madison, WI, USA) for a 50 µl final reaction volume. The cycling parameters were the same as for the 3' RACE reaction. A second amplification was performed using the nested XL4.R (5'-TTT CCA CTG TTA AGC CAA AGA CC-3'), the 5' nested adapter primer, and 1 µl of the first amplification reaction as template. The reaction mixture and cycling parameters remained the same as for the first amplification. The amplicon was cloned prior to sequencing as described above.

A full-length sequence of the revised sequence of Xla-PXN was PCR amplified with the primer XL7.F (5'-AGA GCT ACT GGG ACA TCC AGT CTA TG-3') using *X. laevis* liver cDNA primed with the lock-docking primer. The reaction mixture and cycling parameters were identical to that described above.

Clones dc23g11 (#5593519) and dc10c11 (#5592223) from which the *X. laevis* liver EST (GenBank Accession BE507877) was obtained, were purchased from the American Type Culture Collection (Manassas, VA, USA). After sequencing the insert ends to confirm their identity, both clone inserts were sequenced by using the transposon-mediated template generation system (TGS) (MJ Research, Waltham, MA, USA), a more rapid method than the directional deletion method.

In parallel to sequencing the purchased clones, PCR amplification of the putative FBP-like contig was performed with *X. laevis* liver cDNA. Amplification was performed using the lock-docking primer and forward primer XL8.F (5'-GAA TGT TTG GCA GAG GAC TG-3') directed to the foremost 5' end of the EST contig. The PCR reaction mixture was: 1X Takara

buffer, 2 mM dNTPs, 0.2 mM each primer, 1.25 U Takara ExTaq polymerase mix (PanVera Corp., Madison, WI, USA) for a 50 µl final reaction volume. The polymerase mix used has proofreading capability. The touch-down PCR cycling parameters were: (1) 94 °C for 2 min; (2) 94 °C , 30 sec; (3) 60 °C, 30 sec minus 0.5 °C per cycle (4) 72 °C , 3 min (repeat steps 2 through 4 for 19 cycles); (5) 94 °C, 15 sec; (6) 50 °C, 15 sec; (7) 72 °C, 3 min (repeat steps 5 through 7 for 24 cycles); (8) fill-in at 72 °C for 5 min. After cloning the resulting amplicon, the insert was sequenced by primer walking.

Design of FBPL degenerate primers

Deduced peptide sequences from the FBP32 homologues identified in the public databases were aligned using Block Maker (blocks.fhcrc.org/blocks/blockmkr/make_blocks.html). Specifically, the FBPL sequences included in the multiple alignment were the two tandem domains of MsaFBP32, the steelhead trout FBPL, completed after 3' extension of the trout liver EST (GenBank Accession T23111), and the published frog liver PXN1_XENLA (GenBank Accession P49263). The alignment produced as output was processed by the CODEHOP program [201] (blocks.fhcrc.org/blocks/codehop.html), using the default parameters, which searches for segments within the alignment without gaps (i.e. blocks) to which hybrid primers are designed to anneal to the reverse translated DNA. Two forward and two reverse primers were selected from the program's output. These are DFBP1.F (5'-CCA CAC CGA GAA GCA GAT GVA YCC NTG GTG-3'), DFBP2.F (5'-AGA GAA TCA ACG GAG CCG ARA THM RNA T-3'), DFBP3.R (5'-AGT TTC CGA TGC GGA TCT CNR CNC CRT T-3') and DFBP4.R (5'-CGG GGA TCA TGA CGG TCA CRT ANY KNC

C-3'). The last letter in the primer's name, F for forward and R for reverse, specifies its orientation.

Detection of FBPLs by MOPAC

PCR amplification of MsaFBP32 homologues from genomic DNA was performed with the following reaction mixture: 1X EnzOne 2000 buffer, 1.5 mM MgCl₂, 0.2 mM dNTPs, 0.5 µM each primer, 1 µg genomic DNA and 2 U of EnzOne 2000 polymerase (ID Labs, London, Ontario, Canada). Note the use of genomic DNA rather than cDNA. Instead of cDNA, genomic DNA was used as a PCR template because it was the most readily available genetic material from non-conventional animal model species. The same cycling conditions used for MOPAC of MsxMcFBP32 in chapter 2 are used.

Sequencing of zebrafish FBPL ESTs

Sex-specific primers Bre10.F (5'-CGG TTT GTT TCT GGC GAA G-3') and Bre11.F (5'-GCA AAC AGT CTC CTC TGT GTT C-3') were designed for confirming by 3' RACE the expression of an FBPL in gonads. The head kidney library clone (IMAGE: 6960385) was purchased from Open Biosystems (Huntsville, AL, USA) and fully sequenced.

Sequencing of steelhead trout FBPL EST

From the *Onchorhynchus mykiss* liver EST sequence (GenBank Accession T2311) an upstream primer for 3' RACE was designed, Om1.F (5'-AGA CAC AAC GTC CAG TTG ATA AAG TGC CTG AG-3') for 3' RACE. Five µg of rainbow trout liver total RNA was reverse transcribed with the lock-docking primer as described above. The PCR reaction mixture was as follows: 1X EnzOne 2000 buffer, 1.5 mM MgCl₂, 0.2 mM dNTPs, 0.2 mM each primer, and

2 U EnzOne2000 polymerase (ID-Labs), for a final reaction volume of 50 μ l. The cycling parameters are: (1) 94 °C for 2 min; (2) 94 °C , 40 sec; (3) 45 °C, 1 min (4) 72 °C , 2 min (repeat steps 2 through 4 for 29 cycles); (5) fill-in at 72 °C for 5 min.

For 5' RACE, the cDNA was produced from steelhead trout liver total RNA using the Invitrogen GeneRacer kit according to the manufacturer's instructions. A gene specific primer, Omy7.R (5'-TGG ACC TCT ATT CCA CCG GAG AC-3') is used in combination with the kit's anchored primer to amplify the remaining 5' end of the trout FBP-like transcript. As positive control, the 5' end of β -actin transcript was amplified since it is typically an abundantly expressed gene. The β -actin gene specific primer used is OmyACT2.R (5'-TAC AGG GAC AAC ACG GCC TGG ATG G-3'). The PCR reaction mixture was: 1X Takara buffer, 2 mM dNTPs, 0.2 mM each primer, 1.25 U Takara rTaq polymerase (PanVera Corp., Madison, WI, USA) for a 50 μ l final reaction volume. The touch-down PCR cycling parameters were: (1) 94 °C for 2 min; (2) 94 °C , 30 sec; (3) 65 °C, 30 sec minus 0.5 °C per cycle (4) 72 °C , 3 min (repeat steps 2 through 4 for 19 cycles); (5) 94 °C, 15 sec; (6) 55 °C, 15 sec; (7) 72 °C, 3 min (repeat steps 5 through 7 for 24 cycles); (8) fill-in at 72 °C for 5 min.

The full-length transcript of the above gene was PCR amplified using the primer Om11.F (5'-TCT CCC ACA TGG TGT CCT GGA G-3') in the equivalent of a 3' RACE as described above.

All of the amplicons were ligated to plasmid vector by T/A ligation as described above in preparation for sequencing.

Sequencing of fruit fly FBPL EST

The plasmid clone, LP08801, whose 5' EST (GenBank Accession AI295227) presented similarities to the bass lectin domain, was purchased from Research Genetics (Huntsville, AL, USA). This clone's provenance is from a larva, early pupal stage cDNA library prepared by the Berkeley *D. melanogaster* Genome Project. Sequencing of the insert was performed by directional deletion method using an ExoIII/S1 Deletion kit (Fermentas; Hanover, MD, USA). The plasmid (50 µg) was first linearized by digesting with the restriction enzyme *Bgl*II (25 U) (Promega) for 2 h at 37 °C, then 10 µg treated with Klenow fragment of *E. coli* polymerase I (New England Biolabs, Beverly, MA, USA) for 10 min at 37 °C to fill-in the ends, and phenol/chloroform purified to remove the enzymes. To create an end digestible with *Exo*III/S1 nuclease, the plasmid was digested with 10 U of *Eco*RI (Promega) for 1 h at 37 °C. Directional deletion with *Exo*III/S1 nuclease mix was performed at 37 °C with aliquots containing 5 µg of plasmid sampled every minute. The plasmid of each digest reaction was recircularized with T4 DNA ligase and transformed into chemically competent *E. coli* JM109. Selected plasmids were sequenced as already described using the T7 polymerase site adjacent to the insertion site of the plasmid vector, pOT2a (www.resgen.com/products/DEST.php3). To complete the sequence of the transcript of the cDNA sequenced from the LP08801 clone, a *D. melanogaster* bacteriophage cDNA library prepared from larvae was purchased (Stratagene, La Jolla, CA, USA). A PCR amplification approach was taken using bacteriophage arm-specific primers in combination with a gene-specific primer. Plate lysates of bacteriophage were prepared by inoculating 1 ml ($A_{600}=2$) of overnight cultured *E. coli* Y1090r with 50,000 bacteriophage. After 15 min incubation at 37 °C, the transduced cells were mixed with 7 ml

soft agar and plated on 150 mm Luria-Bertani agar plates. Twenty plates were prepared this way to obtain a library representation of 1,000,000 PFUs. The bacterial lawn was lysed by incubating at 37 °C for 12 h, and the released bacteriophages were suspended in SM buffer. A 200 µl aliquot of the lysate was heated at 100 °C for 10 min to release the DNA.

Confirmation of the presence in the lysates of bacteriophage clones that contain the LP08801 insert was performed by PCR amplification. Primers specific for LP08801 sequence, Dme5.F (5'-TCG TCT TCT CGA ATG CGA ATC TC-3') and Dme2.R (5'-AAA GTT CAG CGG CTG GCG GTT GTA GAA-3'), were used to test the 20 lysates yielding a 674 bp amplicon when positive. The PCR reaction mixture is 1X EnzOne2000 Buffer, 0.2 mM dNTPs, 1.5 mM MgCl₂, 0.2 mM of each primer, and 2 U EnzOne2000 polymerase (ID Labs) and 5 µl of boiled lysate for a final reaction volume of 50 µl. The cycling parameters were: (1) 94 °C for 5 min; (2) 94 °C , 30 sec; (3) 60 °C, 30 sec (4) 72 °C , 1 min (repeat steps 2 through 4 for 34 cycles); (5) fill-in at 72 °C for 1 min. The positive lysates were used for 5' RACE. The PCR reaction consisted of the primers Lgt11.F (5'-GAC TCC TGG AGC CCG-3') and Lgt11.R (5'-GGT AGC GAC CGG CGC-3') specific for the bacteriophage arms and the nested gene-specific primers Dme4.R (5'-TTT GCC AAA GTC CAG ACG CAC CAG-3') and Dme6.R (5'-ACC AGC GCC ATC AGT GGA AAT C-3'). The first PCR reaction mixture was: 1X EnzOne2000 Buffer, 0.2 mM dNTPs, 1.5 mM MgCl₂, 0.2 mM of Lgt11.F or Lgt11.R primer, 0.2 mM Dme4.R, and 2U EnzOne2000 polymerase (ID Labs), and 5 µl of boiled lysate for a final reaction volume of 50 µl. The cycling parameters were: (1) 94 °C for 5 min; (2) 94 °C , 30 sec; (3) 60 °C, 30 sec (4) 72 °C , 1 min (repeat steps 2 through 4 for 34 cycles); (5) fill-in at 72 °C for 1 min. The second nested PCR reaction mixture consisted of: 1X EnzOne2000

Buffer, 0.2 mM dNTPs, 1.5 mM MgCl₂, 0.4 mM of Lgt11.F or Lgt11.R primer, 0.4 mM Dme6.R, and 1U EnzOne2000 polymerase (ID Labs), and 1 µl of 1/100 dilution of the first reaction for a final reaction volume of 25 µl. The resulting amplicons were cloned and sequenced as described above.

Analysis of expression of CG9095 during fruit fly development

One µg each of poly (A)⁺ RNA from embryo, larval, and adult stages of *D. melanogaster* (Clontech Laboratories, Palo Alto, CA, USA) were reverse-transcribed with Thermoscript kit (Invitrogen) at 55 °C, as instructed by the manufacturer. The PCR reaction mixture was: 1X EnzOne 2000 buffer, 1.5 mM MgCl₂, 0.2 mM dNTPs, 0.4 µM each primer, 1 µl of cDNA, 1U EnzOne 2000 polymerase (ID Labs, London, Ontario, Canada) in a final volume of 25 µl. The cycling parameters were as follows: (1) 94 °C, 2 min; (2) 94 °C, 30 sec; (3) 60 °C, 30 sec; (5) 72 °C, 1 min (repeat steps 2 through 5 for 39 cycles). The primers used were Dme5.F (5'-TCG TCT TCT CGA ATG CGA ATC TC-3') and Dme2.R (5'-AAA GTT CAG CGG CTG GCG GTT GTA GAA-3'). Primers directed to a ubiquitous ribosomal protein (rp49) [202], rp49.F (5'-AAG ATC GTG AAG AAG CGC AC-3') and rp49.R (5'-AAC TTC TTG AAT CCG GTG GG-3'), were used as controls of cDNA synthesis completion.

Multiple alignment of peptide sequences

An alignment of the deduced peptide sequences from the identified members of the FBPL family was produced with Clustal_X v.1.81 [203] using default settings. The presentation of the alignments was prepared with GeneDoc v.2.6.002 [204]. Shading of similarities reflects groupings in the BLOSUM62 matrix [205].

Phylogenetic analysis

Distance analysis of polypeptide sequences possessing an FBPL motif was performed using Neighbor-Joining analysis [206] as implemented in Clustal_X v.1.81. Reliability of branching order was estimated using a bootstrap test [207].

Results and Discussion

Meeting the parents

Interest in identifying the parental origin(s) of the MsxMcFBP32 cDNA sequence in the hybrid bass led to the search for homologous sequences in the striped bass (*M. saxatilis*) and white bass (*M. chrysops*). A DNA primer (FBP1.F) specific to the 5' end of the MsxMcFBP32 cDNA was used to amplify the full gene sequence from RNA isolated from liver in each species. Both the striped bass (MsaFBP32) and the white bass (MchFBP32) yielded a 1,144 bp amplicon. To ensure an amplicon with the highest fidelity to the cDNA, a proofreading DNA polymerase mixture was used in the PCR. Increased accuracy of the cDNA sequence was further gained from multiple sequences: 10 clones were sequenced for MsaFBP32 and five clones for MchFBP32. Comparison of the nucleotide sequence from both species presented few changes (12 positions) (fig. 18). Mutations were either silent or resulted in conserved substitutions in the peptide sequence. The encoded polypeptides were 293 residues long, as in the MsxMcFBP32 protein. Surprisingly, the UTR segments of the cDNA, which are considered to diverge rapidly among species presumably due to diminished selective pressure, were almost identical. The calculated biochemical properties for each species lectin were as follows: MsaFBP32 (Mr 32,393.1 Da, pI 6.16) and MchFBP32 (Mr 32,308 Da, pI 6.01). The calculated mass for MsaFBP32 was closer to the mass estimated for the striped bass serum

lectin from mass spectrometry (32,388.2 Da) resolving the discrepancy previously mentioned for MsxMcFBP32. The high nucleotide identity, even within non-coding regions, between the lectin homologues from parental species-specific lectin reflect the close relatedness between these sister taxa [135].

```

>MsFBP CTGGACTCCA GGGATAAAAG ATCTGTTCTA ACCAGGAAGC AGGAATAATG AGGCACAGTG
      G L Q G . K I C S N Q E A G I M R H S V
>McFBP CTGGACTCCA GGGATAAAAG ATCTGTTCTA ACCAGGAAGC AGGAATAATG AGGCACAGTG
      G L Q G . K I C S N Q E A G I M R H S V
#1 -----

>MsFBP TGGTATTCTT GTTGCTGCTC CTCTTAGGGG CGTGTTCAGC TTACAACATAT AAAAATGTGG
      V F L L L L L L L G A C S A Y N Y K N V A
>McFBP TGGTATTCTT GTTGCTGCTC CTCTTAGGGG CGTGTTCAGC TTACAACATAT AAAAATGTGG
      V F L L L L L L L G A C S A Y N Y K N V A
#61 -----

>MsFBP CCTTGCGTGG AAAAGCGACT CAGTCGGCAC GTTATTTGCA CACACATGGA GCCGCCTACA
      L R G K A T Q S A R Y L H T H G A A Y N
>McFBP CCTTGCGTGG AAAAGCGACT CAGTCGGCAC GTTATTTGCA CACACATGGA GCCGCCTTCA
      L R G K A T Q S A R Y L H T H G A A F N
#121 -----*--

>MsFBP ACGCCATTGA TGGAAACCGT AACTCTGACT TCGAAGCTGG ATCGTGCACC CACACTATTG
      A I D G N R N S D F E A G S C T H T I E
>McFBP ACGCCATTGA TGGAAACCGT AACTCTGACT TCGAAGCTGG ATCATGCACC CACACTGTTG
      A I D G N R N S D F E A G S C T H T V E
#181 -----*-----*--

>MsFBP AACAGACCAA CCCCTGGTGG AGAGTGGACC TACTGGAGCC CTACATCGTC ACCTCCATCA
      Q T N P W W R V D L L E P Y I V T S I T
>McFBP AACAGACCAA CCCCTGGTGG AGAGTGGACC TACTGGAGCC CTACATCGTC ACCTCCATCA
      Q T N P W W R V D L L E P Y I V T S I T
#241 -----

>MsFBP CCATCACCAA CAGAGGAGAC TGCTGTCCAG AAAGGCTCAA CGGGGTGGAG ATTCACATCG
      I T N R G D C C P E R L N G V E I H I G
>McFBP CCATCACCAA CAGAGGAGAC TGCTGTCCAG AAAGGCTCGA TGGAGCGGAG ATTCACATCG
      I T N R G D C C P E R L D G A E I H I G
#301 -----*--*--*--

>MsFBP GCAACTCTAT ACAAGAAAAT GGTGTTGCAA ACCCAAGGGT TGGTGTAATT TCTCATATCC
      N S I Q E N G V A N P R V G V I S H I P
>McFBP GCAACTCTTT ACAAGAAAAT GGTGTTGCAA ACCCAAGGGT TGGTGTAATT TCTCATATCC
      N S L Q E N G V A N P R V G V I S H I P
#361 -----*-

>MsFBP CTGCAGGGAT CTCACATACT ATCAGTTTCA CTGAACGTGT GGAGGGACGT TACGTGACTG
      A G I S H T I S F T E R V E G R Y V T V
>McFBP CTGCAGGGAT CTCACATACT ATCAGTTTCA CTGAACGTGT GGAGGGACGT TACGTGACTG
      A G I S H T I S F T E R V E G R Y V T V
#421 -----

>MsFBP TGCTTCTACC TGGTACAAAC AAGGTTCTTA CACTCTGTGA AGTGGAGGTT CATGGGTACC
      L L P G T N K V L T L C E V E V H G Y R
>McFBP TGCTTCTACC TGGTACAAAC AAGGTTCTTA CACTCTGTGA AGTGGAGGTT CATGGGTACC
      L L P G T N K V L T L C E V E V H G Y R
#481 -----

```

```

>MsFBP  GAGCCCCAAC TGGAGAGAAC CTGGCCCTCC GAGGAAAAGC CACACAGTCT TCATTGTTTG
        A P T   G E N   L A L R   G K A   T Q S   S L F E
>McFBP  GAGCCCCAAC TGGAGAGAAC CTGGCCCTCA AAGGAAAAGC CACACAGTCG TCATTGTTTG
        A P T   G E N   L A L K   G K A   T Q S   S L F E
#541    -----*-----*-----

>MsFBP  AATCTGGTAT TGCATATAAT GCCATTGATG GGAATCAAGC CAACAATTGG GAAATGGCCT
        S G I   A Y N   A I D G   N Q A   N N W   E M A S
>McFBP  AATCTGGTAT TGCATATAAT GCCATTGATG GGAATCAAGC CAACAATTGG GAAATGGCCT
        S G I   A Y N   A I D G   N Q A   N N W   E M A S
#601    -----

>MsFBP  CCTGCACTCA CACAAAAAAC ACAATGAACC CCTGGTGGCG AATGGATCTG AGCAAAACCC
        C T H   T K N   T M N P   W W R   M D L   S K T H
>McFBP  CCTGCACTCA CACAAAAAAC ACAATGAACC CCTGGTGGCG AATGGATCTG AGCAAAACCC
        C T H   T K N   T M N P   W W R   M D L   S K T H
#661    -----

>MsFBP  ACAGAGTGTT TTCTGTTAAG GTAACCAACC GAGATTCATT TGAAAAACGA ATCAATGGAG
        R V F   S V K   V T N R   D S F   E K R   I N G A
>McFBP  ACAGAGTGTT TTCTGTTAAG GTAACCAACC GAGATTCATT TGAAAAACGA ATCAATGGAG
        R V F   S V K   V T N R   D S F   E K R   I N G A
#721    -----

>MsFBP  CTGAGATCCG AATTGGAGAT TCCCTCGACA ACAACGGCAA CAACAATCCC AGGTGTGCTG
        E I R   I G D   S L D N   N G N   N N P   R C A V
>McFBP  CTGAGATCCG AATTGGAGAT TCCCTCGACA ACAACGGCAA CAACAATCCC AGGTGTGCTG
        E I R   I G D   S L D N   N G N   N N P   R C A V
#781    -----

>MsFBP  TGATCACAAG CATCCCAGCA GGTGCTTCTA CTGAATTCCA GTGTAACGGG ATGGATGGCC
        I T S   I P A   G A S T   E F Q   C N G   M D G R
>McFBP  TGATCACAAG CATCCCAGCA GGTGCTTCTA CTGAATTCCA GTGTAACGGG ATGGATGGTC
        I T S   I P A   G A S T   E F Q   C N G   M D G R
#841    -----*--

>MsFBP  GCTATGTTAA CATTGTTATC CCTGGAAGAG AAGAGTACCT GACCCTGTGT GAGGTGGAGG
        Y V N   I V I   P G R E   E Y L   T L C   E V E V
>McFBP  GCTATGTTAA CATTGTTATC CCTGGAAGAG AAGAGTACCT GACCCTGTGT GAGGTGGAGG
        Y V N   I V I   P G R E   E Y L   T L C   E V E V
#901    -----

>MsFBP  TGTATGGCTC TGTCTGGAT TAGGTGTCAG TACTAATACT GTTGAATGTA CACAAACAAA
        Y G S   V L D   . V S V   L I L   L N V   H K Q N
>McFBP  TGTATGGCTC TGTCTGGAT TAGGTGTCAG TACTAATACT GTTGAATGTA CGCAAACAAA
        Y G S   V L D   . V S V   L I L   L N V   R K Q N
#961    -----*-----

>MsFBP  ACAAATAGT AGATTAAGCT TTTTGTATTG TTTCCATTCA AAATAAGACA GAGATGGTCT
        K I V   D . A   F L I V   S I Q   N K T   E M V L
>McFBP  ACAAATAGT AGATTAAGCT TTTTGTATTG TTTCCATTCA AAATAAGACA GAGATGGTCT
        K K V   D . A   F L I V   S I Q   N K T   E M V L
#1021   -----*-----

>MsFBP  TATCCAATAA AATTACATCA CGAAAAAAA AAAAAA:: ::::
        S N K   I T S   R K K K   K K ? ?
>McFBP  TATCCAATAA AATTATATCA CGAAATGSAA AAAAAAAA AAAA
        S N K   I I S   R N ? K   K K K K
#1081   -----*-----**+---** ****

```

Figure 18. Pairwise alignment of MsaFBP32 and MchFBP32 full-length cDNAs.

Asterisks (*) in consensus line at the bottom indicate nucleotide differences, dashes (-) are identities, and colons (:) are gaps. The deduced peptide sequence is presented below its corresponding cDNA. Nucleotide position numbering is listed in the left margin. MsFBP32: striped bass lectin, McFBP32: white bass lectin

To confirm the length of the cDNA, a Northern analysis on liver RNA was performed. A similar length transcript of the expected length (1.2 kb) (fig. 19) hybridized to the cDNA probe, although a larger but weaker hybridizing species was also observed above it. The higher band hybridized more strongly appeared more brightly when hybridizing to poly(A)⁺-selected RNA, corroborating the faint band observed in the lane of total RNA. This band possibly represents either an alternative transcript, or a different transcript sharing the FBPL motif.

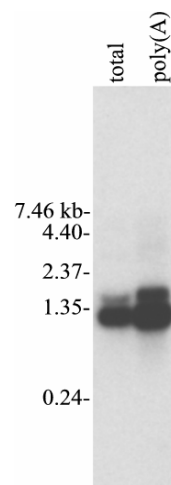


Figure 19. Northern analysis of striped bass liver RNA probed with FBP32 probe.
Total, total RNA; poly(A), mRNA

RNA extracted from various organs was also probed by Northern analysis to test if the lectin was expressed extrahepatically (fig. 20). Only a faint band of corresponding size was observed from intestine, and a larger faint smear appeared for gill tissues. Thus, it appears that the liver, the principal producer of blood proteins, is the major site of this lectin's synthesis, although possibly not the only one. That it is the only site is uncertain since only a sparse sampling of organs was tested.

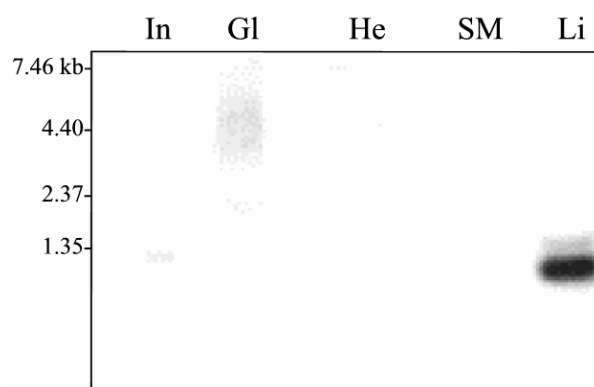


Figure 20. Northern blot detection of tissue expression of MsaFBP32.
In, intestine; Gl, gills; He, heart; SM, skeletal muscle; Li, liver.

A unique lectin family

Comparison to previously described lectin sequence motifs [31, 32] did not reveal any similarity with the bass lectin polypeptide sequence. Further, a search of motif databases including protein profiles [208] (www.expasy.ch/motifscan) produced no significant matches. Therefore, it appeared that the bass lectin was a unique lectin motif of yet unknown ancestry. Although no matches to lectins were identified, a stretch of N-terminus sequence from a single protein named PXN1-XENLA [209] shared significant similarity ($E=8 \times 10^{-26}$) to the bass lectin motif. Surprisingly, PXN1-XENLA, which is described as a pentraxin-fusion protein cloned from the liver of the African clawed frog (*X. laevis*) consists of a fusion domain (i.e. FBP32-like domain (FBPL) linked to a pentraxin domain (fig. 21, Panel A) [25], an unrelated domain that also exhibits lectin activity [210]. PXN1-XENLA along with other discovered mosaic forms of pentraxins [211, 212] are referred to as “long pentraxins” [213], to distinguish them from the prototypic short forms lacking extensions. Interestingly, these long pentraxin domain extensions do not share similarity to each other or to other known domains. In the case of PTX3 from mouse [214], the extension stabilizes multimers through formation of disulfide bridges [215]. In contrast, for PXN1-XENLA the similarity between its fusion domain and

(fig. 22. Panel A). Inspection of the consensus segment indicated that the signal sequence reported for PXN1-XENLA was in fact the end of a FBP-domain. The originally reported PXN1-XENLA cDNA failed to identify this upstream FBP-domain, and this may have been due to two reasons. First, due to an erroneous nucleotide insertion and substitution, the reading frame was shifted and a stop codon was introduced upstream of the starting Met of the putative 5'UTR, respectively (fig. 22. Panel B). Second, the report of the cloning of PXN1-XENLA cDNA [209] remarkably indicates that only 1,739 bp were sequenced from of a 9.5 kb insert of a bacteriophage cDNA clone (XL-PXN1), and erroneously concluded that the deduced polypeptide was complete. To test whether the EST genuinely represented PXN1-XENLA, a 3' RACE reaction was performed using a forward primer (XL1.F) complementary to the portion of the BE508834 not shared with L19881. The amplicon had a length of 1,890 bp, which is the expected length of L19881 including the EST. Finally, sequencing the amplicon confirmed that BE508834 was part of the PXN1-XENLA cDNA.

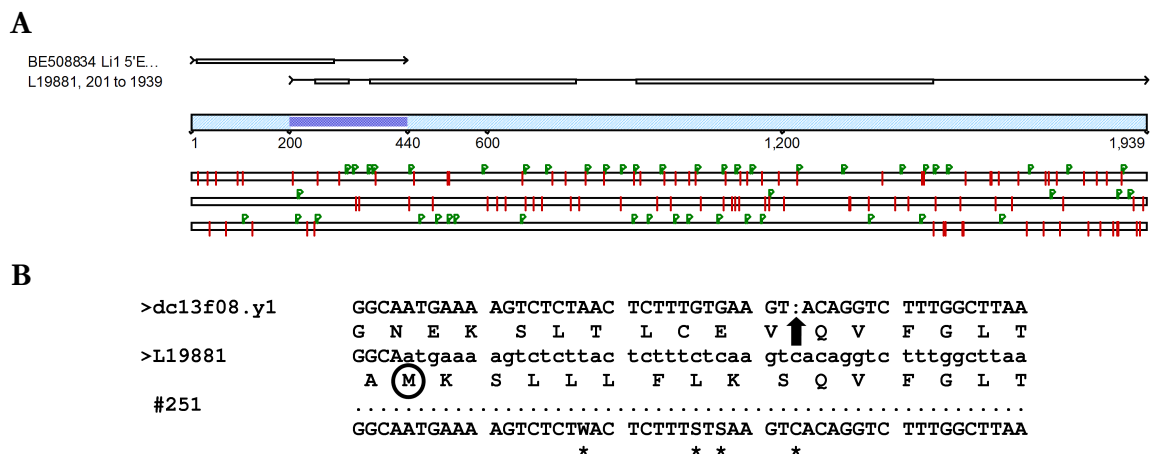


Figure 22. *X. laevis* liver EST matching the 5' end of published PXN1-XENLA cDNA (L19881). **A.** Position of liver EST (GeneBank Accession BE508834) aligned consensus relative to the full cDNA of L19881. Boxed segment of the upper arrow is the incomplete FBP domain, which overlaps with the signal sequence boxed in the lower published sequence. **B.** The appearance of a signal sequence (circled Met) in the original sequence, L19881, coincides with a frameshift caused by an insertion (arrow) that disrupts the FBPL visible in the EST's (IMAGE clone dc13f08.y1) translation. Asterisks below the consensus sequence line mark nucleotide differences.

If the cDNA was 9.5 kb as originally reported, a substantial amount of sequence remained to be determined. 5' RACE amplification produced an amplicon of 2.2 kb (fig. 23). Surprisingly, upon sequencing the amplicon, it became apparent that five FBPLs preceded the pentraxin domain, and not just one FBPL as originally reported (See appendix).

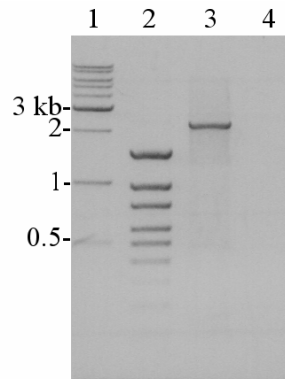


Figure 23. 5' RACE of XL-PXN1.

Lanes: 1) 1 kb marker, 2) 100 bp marker, 3) GeneRacer nested/XL4.R, 4) GeneRacer nested/XL4.R (negative control)

To confirm that this sequence represented an authentic transcript and not a chimera, an artifact frequently present in cDNA libraries, the full-length cDNA was amplified using a forward primer (XL7.F) complementary to the 5' end of the 5' RACE amplicon (fig. 24). The amplicon produced was (3,840 bp) of the size expected from the contig assembled with the 5' RACE amplicon and the 3' RACE amplicon (Fig 25).

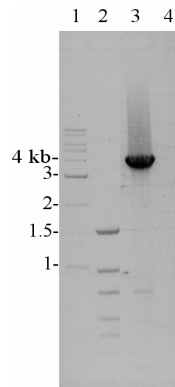


Figure 24. PCR amplification of full length Xla-PXN.
Lanes 1) 1 Kb marker, 2) 100 bp marker, 3) XL7.F/MCSRACE, 4) XL7.F/MCSRACE (negative control)

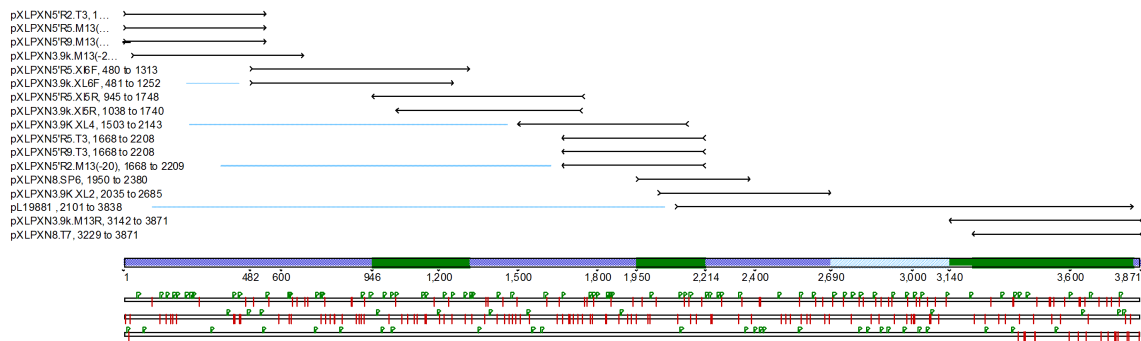


Figure 25. Sequencing contig of full length Xla-PXN.

Although the sequence has been extended to almost twice the length originally reported, it is still shorter than the 9.5 kb of the original cDNA clone. The conciseness of the 5'UTR does support that the sequence is finished. An attempt to extend the sequence searching dbEST identified several 5' ESTs from *X. laevis* liver that matched the 5' end of the revised XL-PXN1 (fig. 26). However, none of these sequences extended further upstream, which would support that the sequence was indeed complete. Consensus found for liver ESTs to the internal sequence of Xla-PXN also lend further support to the veracity of the revised sequence. In summary, the revised sequence (Xla-PXN) links five tandem FBPL domains to a pentraxin domain resulting in a much larger protein (120 kDa).

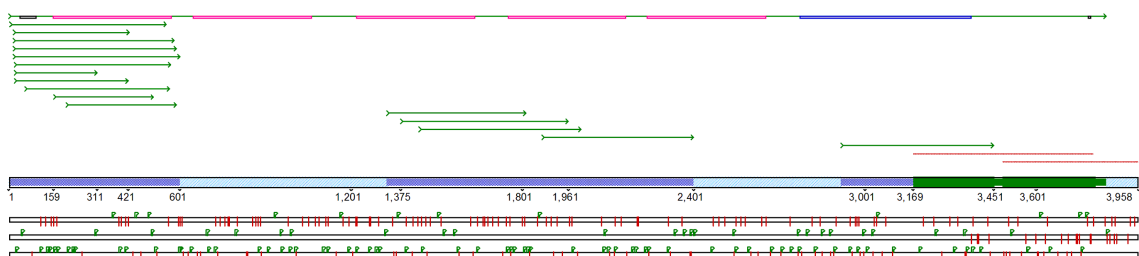


Figure 26. Support for completed ORF of Xla-PXN

Independent ESTs from a *X. laevis* liver library Li1 match the 5' end of Xla-PXN. The top line is the complete CDS after extension by 5' RACE. The first five boxes of this line are the lectin domain and the sixth box is the pentraxin domain.

The march of FBPLs in tetrapods

The search in dbEST identified several *X. laevis* liver ESTs encoding FBPLs that did not match Xla-PXN. A preliminary contig was assembled from these sequences and was extended by searching for ESTs with nucleotide consensus. This *in silico*-assembled cDNA apparently encoded a polypeptide with two tandemly arrayed FBPLs, similar to MsaFBP32 (fig. 27).

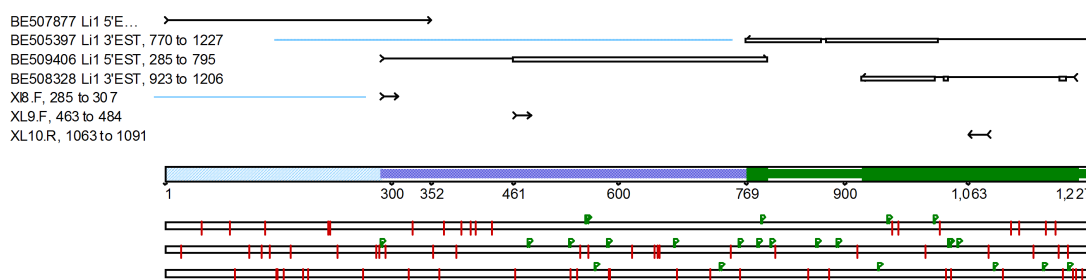


Figure 27. *In silico* assembly *X. laevis* liver EST contig unique from XI-PXN.

Boxes symbolize FBPLs and smaller arrows are the primers designed for PCR amplification to confirm authenticity.

A flaw observed in the sequence assembly is that the overlap of the 5' EST (GenBank Accession BE509406) to the 3' EST (GenBank Accession BE505397) was of only 25 nucleotides, raising doubts about the authenticity of this contig. To verify the sequence, the respective plasmid clones from which these EST had been sequenced were purchased from the ATCC. Both clones were sequenced almost to completion using transposon-mediated gene walking but presented problems. Surprisingly, the plasmid (IMAGE clone 3396692) that

produced EST BE509409 (fig. 28, Panel A) coded for three FBPLs instead of two but the ORF was truncated (i.e. start Met missing). IMAGE clone 3397988, the source of EST 507877, contained a composite transposon previously described in *X. laevis* [218], that interrupts the lectin ORF, as well as causing other interruptions illustrated in the 3rd reading frame (fig. 28, Panel B). Such problems have become evident with the advent of high-throughput sequencing of cDNA libraries [219, 220]. Analysis of human EST sequences indicated that formation of chimeras in cDNA clones might be as high as 11% [219]. What was clear is that this transcript encoded more than two FBPL domains, which are not joined to a pentraxin domain.



Figure 28. Partial sequencing path of purchased *X. laevis* liver cDNA clones.
A. clone 3396692, B. clone 3397988

Finally, to obtain the authentic cDNA sequence, liver RNA was amplified by 3' RACE with a primer complementary to the 5'EST (fig. 29). The sequence of the 2.2 kb amplicon generated (fig. 30) contained a 1,826 bp ORF that encoded a 63.3 kDa polypeptide with four tandem FBPLs (XlaII-FBPL). Inclusion of what is judged authentic 5' end sequence from the chimeric clone (IMAGE clone 3397988) extends the sequence upstream to a final length of 2,548 bp, without affecting the described ORF (See Appendix).

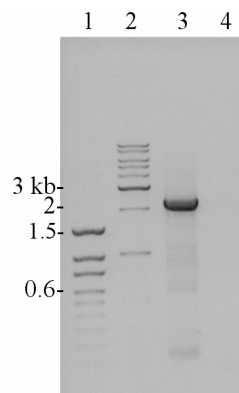


Figure 29. PCR amplification of full-length XlaII-FBPL. Lanes 1) 100 bp marker, 2) 1 kb marker, 3) XL8.F/MCSRACE, 4) XL8.F/MCSRACE (negative control). Gel: 1.2% agarose, 1X TAE

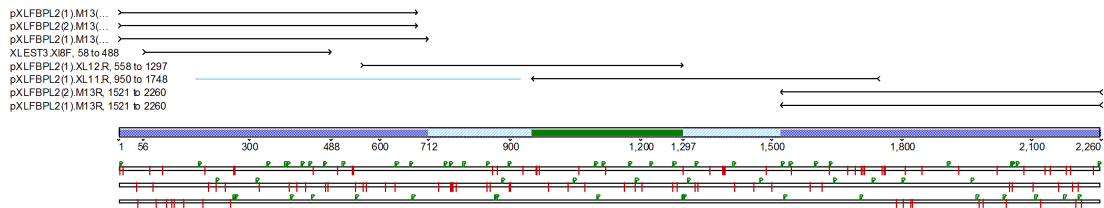


Figure 30. Sequencing contig of the liver XlaII-FBPL containing four FBPL domains

This finding indicates that in *X. laevis*, like in bass, the liver is a site of synthesis of multiple FBP domain-containing proteins. Differences in the number of FBP domains in both species suggest that they are unlikely to be orthologous. Interestingly, the wide range in domain number presented by these sequences illustrates an apparent lack of structural restrictions for the formation of higher order arrangements.

A unique result from the dbEST search from *X. laevis* cDNA libraries is the detection of transcripts containing FBP-like domains expressed early in development (i.e. neurula stage 19-23) which is the first evidence that FBPL play a role during development. Because none of these sequences was verified by cloning, only those assemblies that contained multiple sequences providing added nucleotide coverage were included here. Two distinct putative transcripts were assembled from embryonic stage libraries, constructed from different embryonic stages (i.e. neurula stage 13 through 48). The first contig (fig. 31, 3rd reading frame) encodes what apparently is a single FBPL. The second contig (fig. 32, 1st reading frame) also encodes for a single FBPL, but is different from the first.

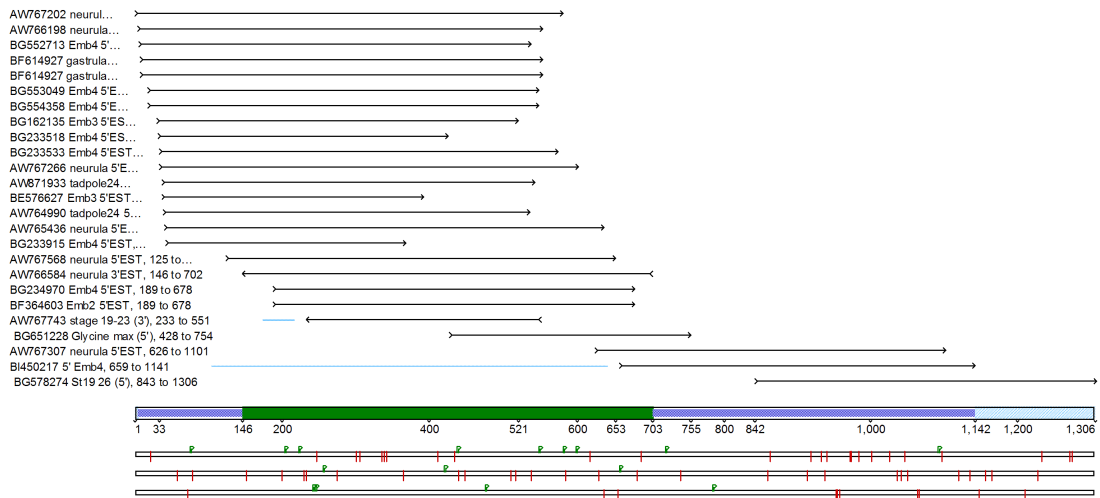


Figure 31. Multiple alignment of *X. laevis* neurula ESTs encoding single FBPL (XlaEST2)

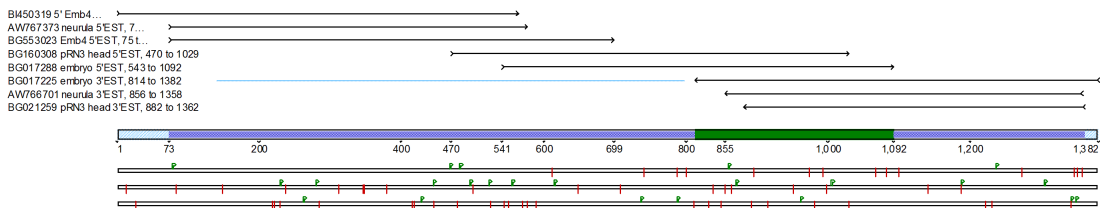


Figure 32. *In silico* assembled contig (Xla-neurula) from *X. laevis* embryos encoding a single FBPL. (GenBank Accession AW767373, BG553023, BG160308, BG017288, BG017225, AW766701 and BG021259)

Both ORFs resulting from these assemblies appear complete, since they contain a starting Met followed by a signal peptide, as predicted by SignalP. Expression of these transcripts correlates with presence of terminally-fucosylated, O-linked glycoconjugates on primordial germ cells (PGC) [221] so it is conceivable that these lectins may play a role in cell adhesion during PGC migration.

The genetic tractability of the diploid *Xenopus tropicalis* has renewed interest in its potential as a model for studying late embryonic development [222]. This in turn has led to an increased effort to develop publicly accessible genomic tools. The recently announced (genome.jgi-psf.org/xenopus1/xenopus1.info.html) plan for sequencing genome of *X. tropicalis* will undoubtedly provide an invaluable resource for annotating the multiple genes that contain FBP-like domains. Continuing the data mining approach, two homologues from *X. tropicalis* were identified from the dbEST. Both assemblies originate from neurula-stage cDNA libraries again supporting the significance of FBPs in development. The first assembly (fig. 33) resulted in a 1,115 bp cDNA encoding a 295 residue mature polypeptide, designated XtrII-FBPL (Appendix). In contrast to the transcript present in *X. laevis* neurula, this ORF contained two FBPLs, which is the first example of a bass FBP32-like binary FBP identified in the tetrapod lineage.

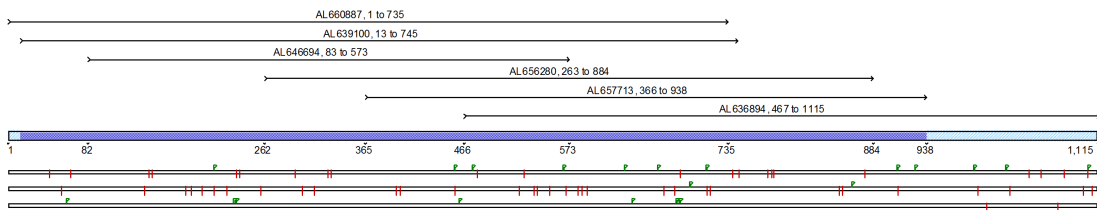


Figure 33. *In silico* assembled contig of XtrII-FBPL.
(GenBank Accession AL660887, AL639100, AL646694, AL656280, AL657713 and AL636894).

The second assembled 1,638 bp cDNA (fig. 34) encodes a 432 residue protein containing three FBPLs, and was designated XtrIII-FBPL (see Appendix).

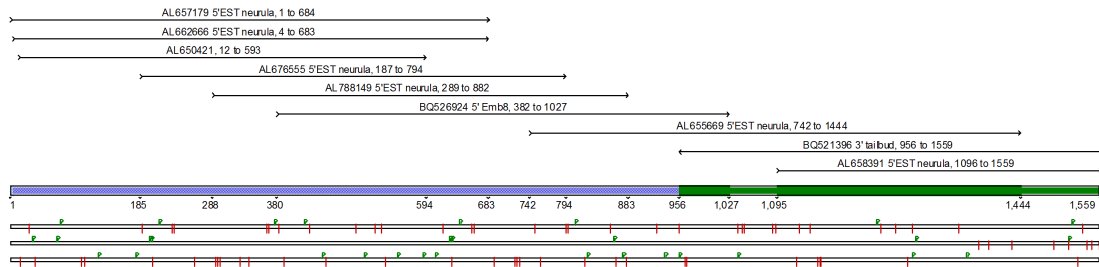


Figure 34. *In silico* assembled contig of XtrIII-FBPL.
(GenBank Accession AL657179, AL662666, AL676555, AL655669 and AL6558391)

The authenticity of both assemblies is supported by coverage from multiple sequences from independent clones. Curiously, neither of these transcripts matches any of the *X. laevis* ESTs deposited so far, and it remains unclear if they are present in this species. Although one assembly from *X. laevis* embryos encodes for three FBPLs, the limited coverage and quality of the sequence makes it premature to assign an ORF with certainty.

The diverse FBPL gene repertoire uncovered in *Xenopus* spp. suggests this domain fulfills diverse roles. The multiple-order concatenates of FBPLs and the pentraxin fusion-protein exemplifies the architectural diversification this domain has undergone. If each domain presumably represents a binding site, concatenation could increase the lectin's avidity through the cluster effect [223]. The diversity of expression sites also argues that FBPLs are not always destined for circulation; ESTs from kidney, heart, and spleen, which contrary to the liver, are not major producers of serum proteins, likely represent FBPLs performing more localized functions. Completion of the sequence for these extrahepatically-expressed FBPL ESTs will tell if they also present unique domain architectures. Members of the genus *Xenopus* are ancient frogs (Suborder Archaeobatrachia) that rarely leave water, however most living frog

species are “modern” frogs (Suborder Neobatrachia) [224] and mainly inhabit land.

Interestingly, some modern frogs have evolved direct development, whereby metamorphosis occurs within the egg enabling these species to bypass the need to deposit eggs in a body of water. Admittedly, there is no evidence correlating the presence of FBPLs with a water-dwelling lifestyle, but the extensive divergence observed in anurans should facilitate investigating this question. Clearly, FBPLs have extensively diversified in the African-clawed frog, a descendant of the lobe-finned fish (Sarcopterygii), but it is presently unknown if this reflects the case within the ray-finned fish. As ray-finned fish frequently possess twice the number of genes present in mammals, evidently the result of a genome-scale duplication that occurred after the two lineages diverged [109], it is likely they have an even greater diversity of FBPLs.

Diversity of FBPLs from ray-finned fish

Undoubtedly, the increasing availability of EST and genomic databases for non-conventional model organisms greatly enhances the possibilities of success in the search of FBPL homologues. However, to trace their ancestry confidently a greater taxonomic diversity is required. Prior to the escalation in EST sequencing that facilitated the discoveries described above, the only homologue detected in GenBank was PXN1-XENLA. The success of the MOPAC procedure with the bass FBP32 suggested that rapid identification of FBPLs could be accomplished by PCR. A trout liver EST, the first FBPL identified in dbEST, (GenBank Accession T23111) provided additional sequence diversity in the multiple sequence alignment. The design of degenerate primers was facilitated by the CODEHOP program [201], which automates the process of identifying those blocks of conserved amino acids optimal for designing primers with limited degeneracy. The alignment provided to the program included

both domains from the bass lectin, the single FBPL from the unrevised PXN1-XENLA and the single FBPL discovered after 3' RACE of the trout EST. From the high similarity blocks identified within the FBPL sequence alignment, CODEHOP returned a number of possible DNA primers in both orientations. A selection of four primers, two forward and two reverse primers, was made after comparing each with the nucleotide alignment of the FBPL sequences to confirm complementarity (fig. 35).

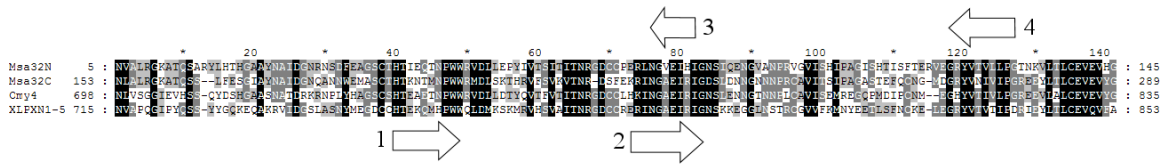


Figure 35. Design of degenerate DNA primers for MOPAC using CODEHOP. Number next to arrow follows designation in Methods and Materials. Msa32N: MsaFBP32 N-FBPL; Msa32C: MsaFBP32 C-FBPL; Omy4: *O. mykiss* FBPL-4; XLPXN1: *X. laevis* XENLA-PXN.

The first attempt at MOPAC using this primer set was performed with liver cDNA from zebrafish, channel catfish (*Ictalurus punctatus*), tilapia (*Oreochromis mossambicus*), and *X. laevis*, as a positive control. Amplification with universal primers of β -actin served as a second positive control for reverse transcription. Expected bands of 137 bp (DFBP1.F/DFBP3.R) and 265 bp (DFBP1.F/DFBP4.R) were only observed in the *X. laevis* control (data not shown). The catfish produced a strongly staining but larger than expected band (650 bp) that was cloned and sequenced. The deduced polypeptide revealed it was unrelated to FBP. The other species produced several weakly staining bands also greater in size than the expected, and sequencing was not pursued. Since liver is not the sole site of FBPL synthesis, a second attempt was made using genomic DNA as template. Only zebrafish genomic DNA was used at first including striped bass genomic DNA as positive control. A band of the expected size (i.e. 137 bp) was produced from zebrafish (fig. 36) as well as for the striped bass (data not shown). Sequencing this band confirmed that it was indeed a segment of a zebrafish FBPL gene.

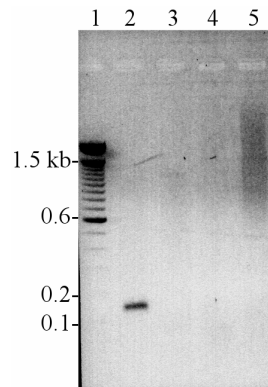


Figure 36. PCR amplification from genomic DNA of a zebrafish FBPL homologue by MOPAC. Lanes 1) 100 bp marker, 2) primers DFBP1.F/DFBP3.R, 3) DFBP1.F/DFBP4.R, 4) DFBP1.F/DFBP3.R, 5) DFBP2.F/DFBP3.R. Gel: 1.2% agarose, 1X TAE

The full-length sequence was obtained by 3' RACE (fig. 37) and by 5' RACE (fig. 38) using total RNA extracted from whole zebrafish. The appearance of two bands from 5' RACE resulted from the amplification of a 5' end truncated form despite the specificity of the technique for amplifying only from full-length transcripts.

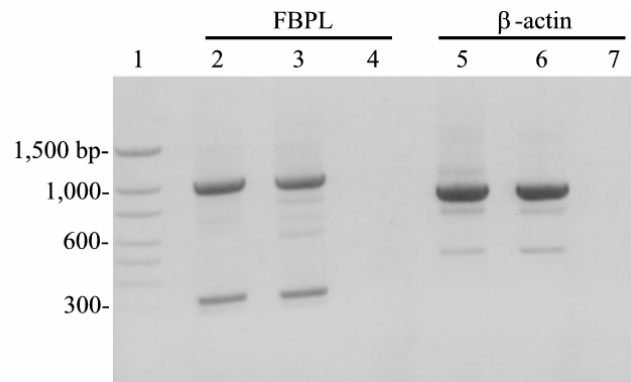


Figure 37. Nested 3' RACE of DreI-FBPL. Lanes 1) 100 bp marker, 2) Bre2.F/MCSRACE (neat), 3) Bre2.F/MCSRACE (1/100), 4) Bre2.F/MCSRACE (negative control), 5) OmyACT.F/ MCSRACE (neat), 6) OmyACT.F/ MCSRACE (1/100), 7) OmyACT.F/ MCSRACE (negative control). Gel: 1.2% agarose, 1X TAE

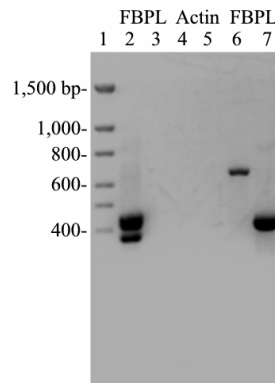


Figure 38. Nested 5' RLM-RACE of DreI-FBPL.

Lanes 1) 100 bp marker, 2) Inner/Bre4.R (nested PCR), 3) Inner/Bre4.R (primer negative control), 4) Outer/OmyACT4R (actin RACE control), 5) Outer/OmyACT4.R (primer negative control), 5) OmyACT1.F/OmyACT4.R (GSP positive control), 6) Bre3.F/ Bre6.R (GSP positive control). Gel: 2 % agarose, 1X TAE

Like the bass lectin, the ORF for the zebrafish homologue (DreI-FBPL) encoded two FBPL domains (Appendix) to which the amplicon obtained by MOPAC mapped to the N-terminal domain (fig. 39).

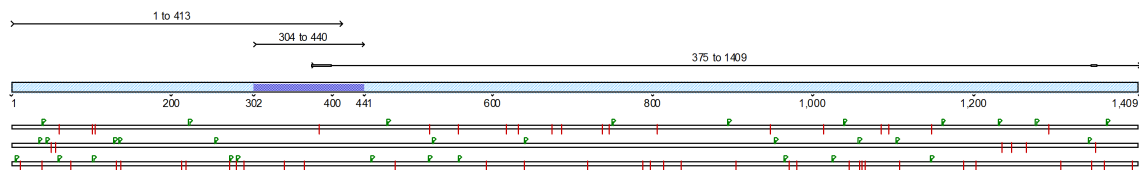


Figure 39. Completion of DreI-FBPL sequence.

The first line is the 5' RACE amplicon, the second line is the MOPAC amplicon and the third line is the 3' RACE amplicon.

Unlike the bass lectin, however, DreI-FBPL is a longer polypeptide (345 a.a.) due to an extension of the COOH-terminal end. Importantly, this segment displays two Cys in addition to the positions conserved in the other FBPLs, suggesting that unique disulfide bridges are likely established in this protein. The universality of this degenerate primer set for amplifying FBPLs was not tested but the success in cloning DreI-FBPL suggests it may constitute a useful tool for identifying FBPLs in species for which gene libraries have not yet been developed.

Cloning of novel gene through MOPAC is advantageous for its speed but it should not substitute for traditional gene library screening since it may miss divergent homologues. This

was apparent during screening of a bacteriophage genomic DNA library (Chp. 4) using the DreI-FBPL cDNA as a probe. Upon sequencing, a positive clone for a second FBPL-like gene was identified. From the partial sequence of a genomic clone, λ Zf13, a complete FBPL was found which was similar but not identical in DNA sequence to the FBPLs of DreI-FBPL. To obtain more sequence downstream, 3' RACE was performed from whole body cDNA (fig. 40). Only after a “nested” PCR on the first amplification reaction did an amplicon become visible (fig. 40, lane 6).

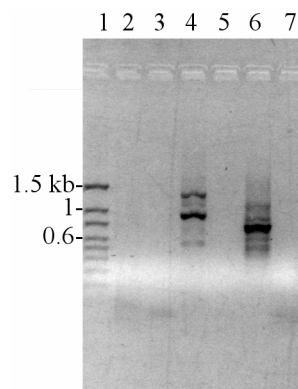


Figure 40. 3' RACE with nested DNA primers of DreII-FBPL identified from λ Zf13. Lanes 1) 100 bp marker, 2) BreIIA.F/MCSRACE (primary PCR), 3) BreIIA.F/MCSRACE (primer negative control), 4) OmyACT1.F/MCSRACE (actin RACE control), 5) OmyACT1.F/MCSRACE (primer negative control), 6) BreIIB.F/MCSRACE (nested PCR), 7) BreIIB.F/MCSRACE (primer negative control). Gel: 1.2% agarose, 1X TAE

The deduced polypeptide of this partial cDNA (DreII-FBPL) contained a single FBPL along with an extension although slightly shorter than observed for DreI-FBPL (see Appendix).

Attempts to complete the sequence by 5' RACE were unsuccessful and no 5' ESTs matched the partial cDNA of DreII-FBPL, and although it remains to be rigorously determined whether the full-length sequence only contains a single FBPL, this possibility is likely since the extensive upstream sequence (7 kb) in the genomic clone shows no evidence of other FBPL domains.

One interesting characteristic that has emerged from screening for FBPLs is the diversity in sites of expression. This varied localization may drive the selection of the multiple FBPLs and points to a subfunctionalization towards tissue-specific operation. The expression of an FBPL in the gonads rather than the liver of zebrafish are one such an example. Two ESTs were initially recognized as containing an FBPL motif: one originated from an ovary cDNA library (GenBank Accession BI983962), and the other from a testis cDNA library (GenBank Accession BQ074381). Alignment of both sequences presented a long segment of consensus but the 5' ends containing the 5'UTR differed. Specifically, the testis EST 5'UTR was shorter than the ovary EST. Two forward primers were designed as specific for each EST to confirm their uniqueness. Using these primers in a 3' RACE with whole-body cDNA from fish of both sexes produced several amplicons (fig. 41). Unexpectedly, the testis-specific primer (Bre11.F) produced the longest amplicon. After sequencing the longest amplicon generated by each primer it was confirmed that they were identical since the amplicon produced by the ovary-specific primer is nested in the larger testis-specific amplicon.

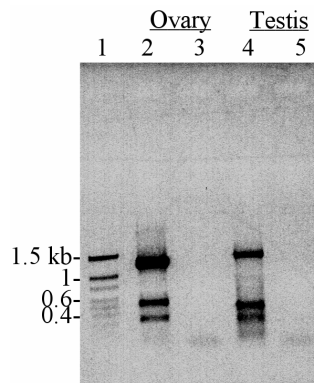


Figure 41. PCR Amplification of DreIII-FBPL with gonadal EST-specific primers. Lanes 1) 100 bp marker, 2) Bre10.F/MCSRACE (primary PCR), 3) Bre10.F/MCSRACE (primer negative control), 4) Bre11.F/MCSRACE, 5) Bre11.F/MCSRACE (primer negative control). Gel: 0.8% agarose, 1X TAE

Alignment of the sequences revealed that the reported testis EST was shorter because of a deleted segment in the 5' UTR. Fortunately, the Bre11.F primer did not span the segment deleted, and was able to amplify the longer cDNA. Analysis of the full cDNA sequence, (DreIII-FBPL) of this gene reveals two FBPLs in tandem like DreII-FBPL, but lacking the extension (See Appendix). The absence of this characteristic makes it more alike to MsaFBP32 than to the other two zebrafish FBPLs. That the above-mentioned transcripts are specific to one sex or the other appears doubtful, and is most likely that they present in both. The function(s) of this protein in gonadal physiology [225] is unknown, but it is conceivable that it plays a role in phagocytosis during follicular atresia [226] and spermatogenesis [227]. It is of note that a Fuc-binding lectin has been isolated from ova of the European sea bass (*Dicentrarchus labrax*) [228], a species closely related to American moronids [135], but it remains unknown if it represents an FBPL. A fourth FBPL (DreIV-FBPL) was reconstructed (Table 3) from an assembled whole genome shotgun contig (Ensembl NA11125), but to date no matching EST has been deposited demonstrating its expression. Another FBPL apparently expressed outside the liver was detected in a head kidney-derived library (GenBank Accession CD758445). Sequencing (DreV-FBPL) of the clone also revealed a tandem domain arrangement (See Appendix), with no extension. The emerging scenario in zebrafish of multiple, binary domain-FBPLs limited in spatial expression diverge from the observations in *Xenopus* spp. where domain rearrangement apparently predominates rather than duplication.

Table 4. Predicted intron splice sites for DreIV-FBPL gene
(Ensembl Zebrafish Contig NA11125)

Putative exons	cDNA domain encoded	Exon size (bp)	Donor Site (exon intron)	Intron Phase	Acceptor Site (intron exon)	Intron size (bp)
1	N-FBPL	294	ACT GTAAGGAAACCATAGAG	2	TTGTTAATTGCAGCTAG AAA	86
2	N-FBPL	155	ATG GTGAGAATATAACCAT	1	TTGTTGTTTCTGTGCAG GGC	1376
3	C-FBPL	283	CAG GTAACAAAACGTGAGCTC	2	CTCTTCCGTTGATTTT AG GTG	>729 (Gap)
4	C-FBPL	231			Stop codon	

Fish can count

From the deduced FBPLs found in *Xenopus* spp. it is clear that this domain is capable of forming tandem arrays of multiple FBP domains. In teleost fish, however, none of the FBPLs detected exhibit more than two domains. As previously mentioned, the *in silico* search initially detected an EST containing the 5' end-sequenced liver cDNA clone from the steelhead trout, *Oncorhynchus mykiss* (GenBank Accession T23111). The initial identification of this partial sequence as an FBPL was confirmed by 3' RACE. To complete the ORF, 5' RACE was performed that resulted in the amplification of two closely sized bands greater than 1.5 kb (fig. 42, lane 4). Due to the closeness in size of the amplicons, they were TA-cloned without prior separation, and then individually identified during the colony selection process (fig. 43). Clones containing either a 1.9 kb or a 2.1 kb insert were selected for sequencing.

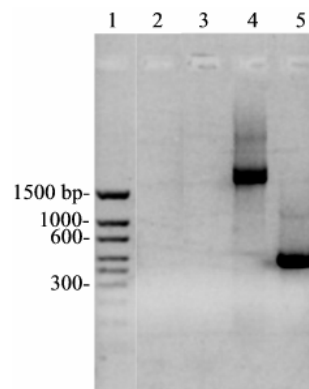


Figure 42. 5' RACE amplicons of trout FBPL from liver total RNA. Lanes 1) 100 bp marker, 2) GeneR5'/OmyACT2.R (primer negative control), 3) GeneR5'/Omy7.R (primer negative control), 4) GeneR5'/Omy7.R, 5) GeneR5'/OmyACT2.R (actin positive control). Gel: 1.2% agarose, 1X TAE. Two amplicons close in size are evident in lane 4.

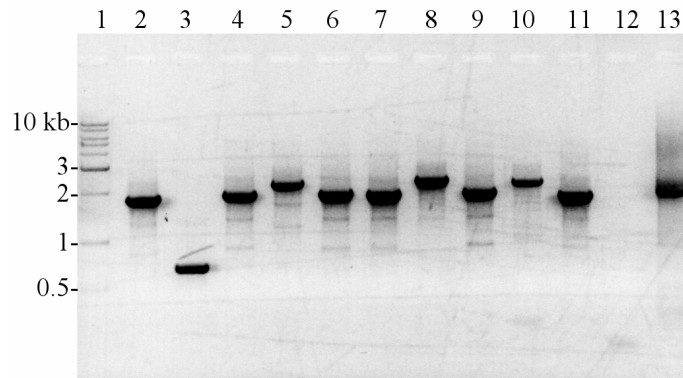


Figure 43. Colony PCR of OmyFBPL4 5' RACE subclones.

Clones two (lane 2), four (lane 4), five (lane 5) and seven (lane 7) were sequenced.

Surprisingly, the longest 5' RACE amplicon (2,960 bp), which extended the length of the initial EST, encoded an ORF (698 aa) with four tandem FBP domains, the first fish FBPL with greater than two domains (fig. 44).

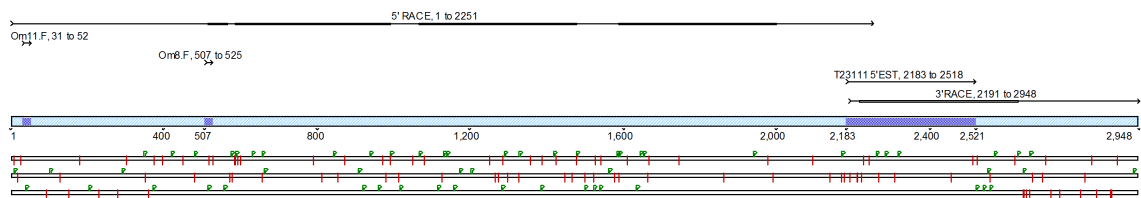


Figure 44. Contig of RACE amplicons for OmyFBPL4.

The arrow labeled T23111 indicates position of EST. First, short box at right indicates the signal peptide-encoding segment. Longer boxes illustrate the four FBPLs. Annealing positions of DNA primers used in PCR amplification of full sequence are illustrated with short arrows.

The sequences revealed that the shorter amplicon had a 118 bp deletion in the 5'UTR, just upstream of the start Met. To examine the possibility that this deletion may be a product of the RACE methodology, a DNA primer, Om11.F, specific for the most 5' end of the 5' RACE amplicon was used in 3' RACE (fig. 44). A second primer, Om8.F, specific for the putative signal peptide, which is downstream from the deletion, was used as a secondary verification. For the first primer, two amplicons were produced (fig. 45, lane 1); the first band was the expected 2.9 kb if no 5'UTR deletion was present but the second band of 1 kb was unexpected. Sequencing of the smaller band indicates it is chimeric since it contained the

5'UTR of OmyFBPL4, including the deletion observed in the 5' RACE amplicon, fused to a microsatellite-like sequence in the 3' half (GenBank Accession AY039629) [229] of the sequence (fig. 46). Because microsatellites are presumably not transcribed, it is possible that contamination of the amplification reaction with genomic DNA generated this amplicon. On the other hand, it may reflect the localization of the microsatellite within the trout FBPL gene leading to an aberrant transcript. This alternative explanation is supported by the localization of the deletion in the smaller 5' RACE amplicon being in the same location as the junction to the microsatellite. The results from both 5' RACE and full length PCR are inconclusive regarding the authenticity of the deleted 5'UTR segment. The elucidation of the structure of this gene will shed light on the possibility of alternative splicing sites that may produce multiple gene products. PCR amplification with the Om8.F primer produced a single 2.5 kb amplicon (fig. 45, lane 4) of the expected size, which supports the authenticity of the four-FBP domains ORF.

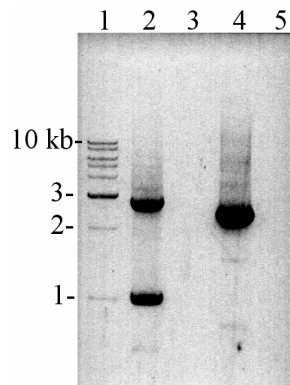


Figure 45. PCR amplification of full-length OmyFBPL4.

Lanes 1) 100 bp marker, 2) Om11.F/MCSRACE, 3) Om11.F/MCSRACE (primer negative control), 4) Om8.F/MCSRACE, 5) Om8.F/MCSRACE (primer negative control). Gel: 1.2% agarose, 1X TAE. Lane 2 uses an upstream primer specific for the most 5' end of the largest band from the 5' RACE. Lane 4 uses a primer specific to the segment missing in the lower band of the 5' RACE.

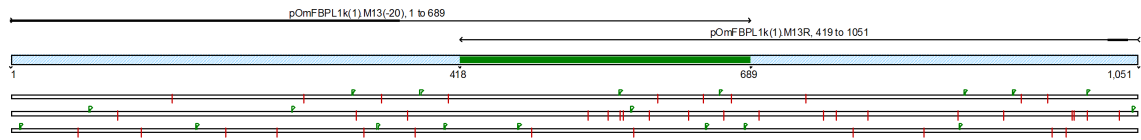


Figure 46. Sequencing contig of 1 kb chimera.

Box illustrates segment that matches the 5'UTR of OmyFBPL4.

The circulation of this lectin in blood has not been confirmed but the presence of a signal peptide suggests the liver secretes it. Many other lectins have been described from blood of the steelhead trout [127, 129], including other unrelated lectins [192] possessing tandem domain arrangements. Pentraxins have also been identified in trout, but do not appear as in *X. laevis* to be fused to an FBPL [122, 230]. From the domain architecture, it is possible that OmyFBPL4 is orthologous to the four domains of XlaII-FBPL.

For some FBPLs identified from EST projects, the length and quality of sequence allowed for assembly of the complete ORFs. Such is the case of a single domain-FBPL from a head-kidney cDNA library of the common carp [231]. Interestingly, this library was produced from cultured head kidney cells treated with LPS and the lymphocyte mitogen concanavalin A, but the impact of these treatments is unknown since differential expression was not tested nor was the library subtracted [232]. Curiously, like zebrafish, this FBPL is expressed in the portion of kidney dedicated to hematopoiesis [233] but unlike zebrafish, it only possesses a single domain. The three-spined stickleback, *Gasterosteus aculeatus*, has become a popular model in the study of morphological evolution due to their rapid diversification [234]. Current efforts in studying the genetic basis for this phenomenon include establishing a genome-wide linkage map, which will permit correlating adaptive phenotypes to the responsible loci [235]. As part of this effort, an EST project was initiated for the production of genetic markers to enhance the resolution of the map (cegs.stanford.edu/index.jsp). Among the latest deposited sequences in dbEST are

two binary FBPLs, CDA91_B12 (GenBank Accession CD508500 and CD508499) and CDA35_C06 (GenBank Accession CD498736 and CD498735) resulting from this EST project. Unfortunately, the libraries were generated from mixed tissues from both sexes, making it difficult to attribute the expression to a specific tissue or sex.

One surprising result came from the purification and cloning of a Japanese eel lectin. As discussed in Chapter 2, the NH₂-terminal peptide sequence of the European eel lectin shares only a few residues with the bass FBP32, and thus they did not initially appear to be related. Searching for serum proteins regulated by changes in water osmolarity, Honda et al. followed a proteomic approach using 2-D electrophoresis to examine plasma of freshwater and saltwater-adapted eels [182]. One particularly abundant protein was cloned from liver and gills that matched the N-terminal reported for the European eel agglutinin [168], which raises the possibility that it was a homologue of the previously characterized fucoselectin from the European eel [236]. Although levels of the Japanese eel FBPL did not correlate with changes in water osmolarity, they appeared to increase upon challenge with bacterial LPS, supporting a role for these lectins in immunity. As the carp head kidney FBPL, only a single FBP domain is present in this protein.

The completion of the genome sequence for the tiger pufferfish, *F. rubripes*, represents a landmark in vertebrate and fish genomics [237], and its release provided the opportunity to search a fish genome for the full repertoire of FBPL genes. Two scaffold assemblies (>100 kb) from the 3rd assembly version by the U.S. Department of Energy's Joint Genome Initiative were identified as containing FBPL domains (for details see Chp. 4). Upon close inspection, each of the two scaffolds apparently encoded a binary FBPL like MsaFBP32. Coincidentally, the draft genome of a different pufferfish species, the green-spotted pufferfish (*T. nigroviridis*)

had also been assembled, allowing comparison between the two species. Similarly, the search of the green-spotted puffer genome identified two scaffolds containing binary FBPL. In comparison to other fish, the pufferfish appears to possess a reduced FBPLs repertoire. A recount of FBPLs uncovered in euteleosts includes singular, binary, and quaternary domain arrangements. This diverse architecture present within these widely diverging fish taxa (i.e. elopomorphs, ostariophysii protacanthopterygians, and acanthopterygians) indicates great divergence of this lectin family along the teleost lineage (fig. 47).

A comparison of these FBPL genes between the two pufferfish species results in clear pairing, which suggests they are orthologous. Interestingly, there is no sign that a quaternary FBPL like that in trout, or a singular FBPL as in the eel and carp. Pufferfish belong to the order Tetraodontiformes, the most derived group of teleosts [238], which are peculiar due to their apparent morphological simplification and genome compaction [239]. Therefore, it appears that spiny-rayed fish (e.g. bass and three-spined stickleback) in general may only possess tandem FBPLs.

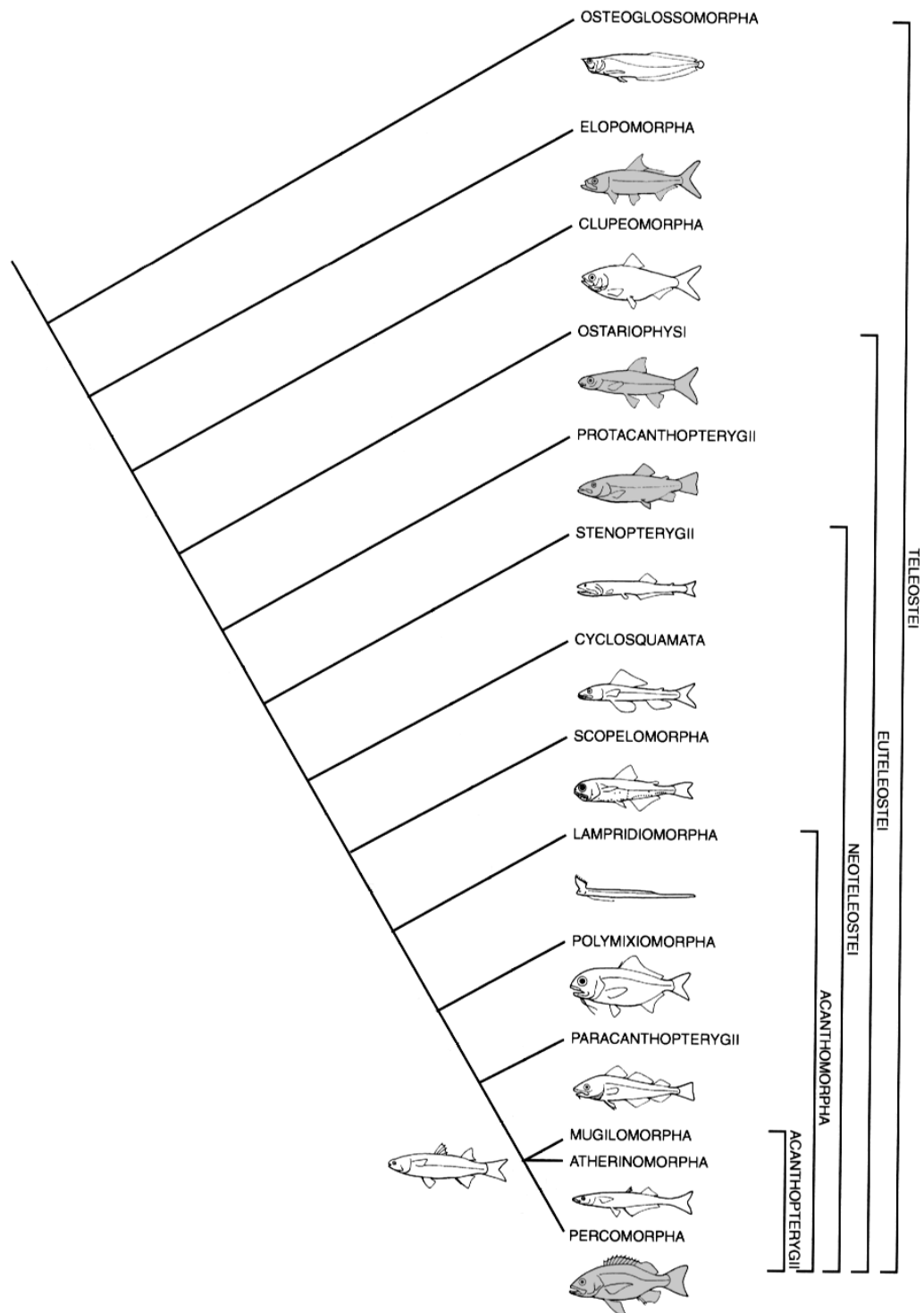


Figure 47. Phylogeny of living fish [238].

An FBPL has been identified in the divisions shaded in gray. Elopomorphs include eels, ostariophysi include zebrafish, protacanthopterygi includes trout and acanthopterygi include sticklebacks, bass and pufferfish. (Adapted from [238]).

FBPLs of invertebrates

Hitherto, only FBPLs from ectothermic vertebrates have been discussed, but their presence in invertebrates was confirmed. During searches of dbEST a partial FBPL motif (GenBank Accession AI295227) was detected from the fruit fly, *D. melanogaster*. The fruit fly has become an invaluable model in the study of innate immunity [240] by facilitating the discovery of the mammalian Toll pathway [241], and therefore revealing a novel family of receptors involved in pathogen recognition [242]. Although TLRs of vertebrates apparently bind pathogen products directly in the fly, Toll interacts with intermediary hemolymph proteins activated by an unknown receptor [17]. The possibility that an FBPL could be this putative receptor was explored by completing the sequence of this interesting EST. To confirm the presence of a full FBPL, the corresponding clone (BDGP/HHMI *D. melanogaster* EST project clone LP08801) was purchased. Sequencing of the plasmid insert (3,201 bp) was completed by directional deletion constructs. Although the FBPL domain was complete, no start Met was present. To complete the ORF, a cDNA library was screened by PCR with a bacteriophage arm-complementary forward primer and an insert specific primer. The resulting 361 bp amplicon included the start Met and a signal peptide predicted with high confidence (fig. 48). Further upstream sequence was added from four complementary ESTs extending the 5' UTR, whereupon several stop codons appeared in frame with the ORF (fig. 49). The completed polypeptide presented the most complex mosaic domain topology so far identified for FBPLs. From the N-terminus, the signal sequence was followed by complement control protein (CCP) domains, also known as sushi domains or short consensus repeats (SCR) [243], the FBPL domain, a C-type lectin-like domain (CTLD), 3 tandem CCPs, and a mucin-like segment [244] that probably is O-glycosylated. Interestingly, past these domains is a segment rich in

hydrophilic residues predicted to form a transmembrane domain [245] (fig. 50). Therefore, LP08801 likely encodes a type I transmembrane receptor with the mucin domain likely serving as a stem to elevate the distal receptor domains away from the cell surface. Only recombinant expression of this protein will confirm its topology with certainty, but the above-mentioned domains are likely extracellular. The designation given to this gene after annotation of the fruit fly genome [246] is CG9095.

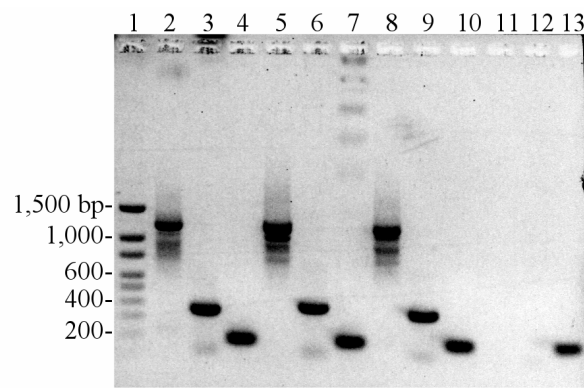


Figure 48. 5' RACE of LP08801 from larval cDNA library lysates.

Lanes 1) 100 bp marker, 2) lgt11.F/Dme6.R, 3) lgt11.R/Dme6.R, 4) Dme5.F/Dme2.R (positive control), 5) lgt11.F/Dme6.R, 6) lgt11.R/Dme6.R, 7) Dme5.F/Dme2.R (positive control), 8) lgt11.F/Dme6.R, 9) lgt11.R/Dme6.R, 10) Dme5.F/Dme2.R(positive control), 11) lgt11.F/Dme6.R (negative control), 12) lgt11.R/Dme6.R (negative control), 13) Dme5.F/Dme2.R (negative control). Gel: 1.2% agarose, 1X TAE

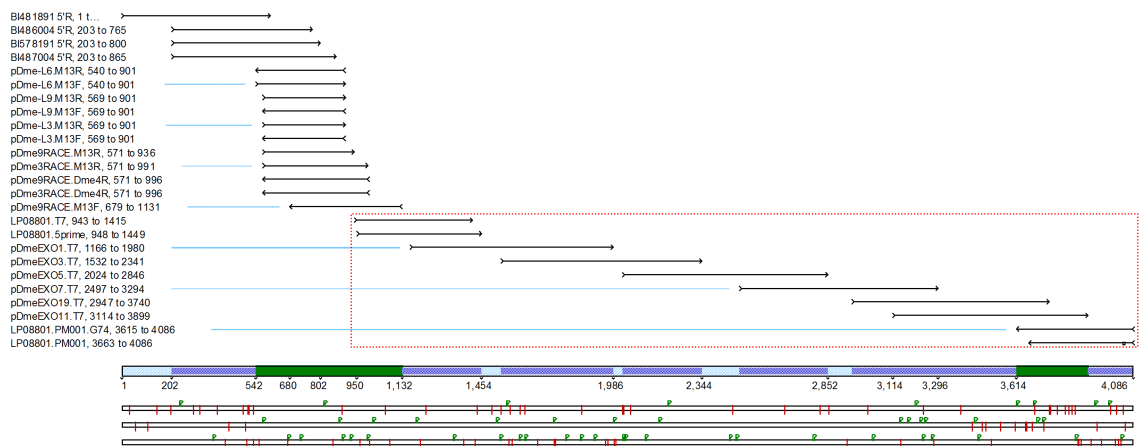


Figure 49. Sequencing contig of the full length *D. melanogaster* CG9095.

Arrows within the rectangle represent the sequence contig for the EST clone LP08801.

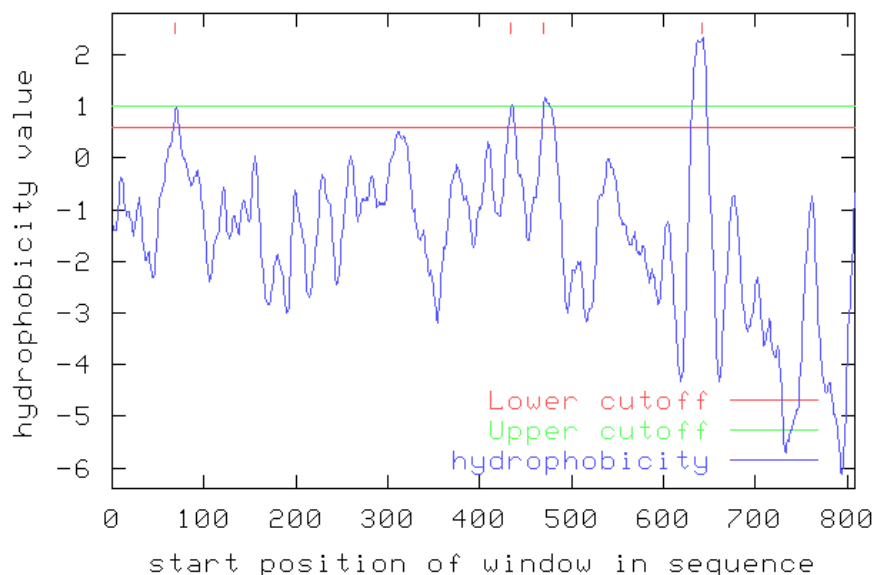


Figure 50. Goldman-Engelman-Steitz hydrophobicity plot of the mature CG9095.

Expression of CG9095 in the larval stage of the fly was evident from the source of the EST but it remained uncertain if it is expressed during other stages of development. Thus, the presence of CG9095 was tested by RT-PCR in embryos and adults. Results indicate that CG9095 is expressed at all stages of fly development, but expression appears higher in the embryo (fig. 51). The continued expression of CG9095 throughout the developmental stages suggests that it derives from a tissue that remains relatively unchanged during metamorphosis, but experimental localization remains to be performed.

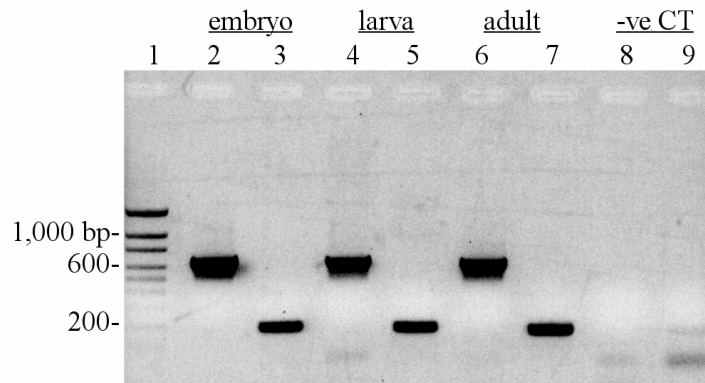


Figure 51. RT-PCR amplification of CG9095 from fruit fly embryonic, larval and adult RNA. Lanes 1) 100 bp marker, 2) Dme5.F/Dme2.R, 3) rp49.F/rp49.R (RT positive control), 4) Dme5.F/Dme2.R, 5) rp49.F/rp49.R, 6) Dme5.F/Dme2.R, 7) rp49.F/rp49.R, 8) rp49.F/rp49.R (negative control), 9) Dme5.F/Dme2.R (negative control). Gel: 1.2% agarose, 1.2X TAE

Searching for proteins likely homologous to CG9095 it was found that the *furrowed* gene [247] shared similar features including the transmembrane domain. Both proteins possess the CTLD followed by CCPs but in the case of *furrowed* there are ten CCPs, in contrast to the three CCPs observed in CG9095. The initial description of *furrowed* cDNA did not include an upstream CCP and FBPL domain despite both being present in the sequenced locus. It appears that these domains were missed because the authors relied on an incomplete composite transcript, built from overlapping truncated cDNA clones, to annotate the exon boundaries. The confirmation of gene annotation of the fruit fly genome [82] through sequencing full cDNA clones has led to a revised *furrowed* cDNA (GenBank Accession BT003214) which encodes for the missing domains. Curiously, a comparison of the LP08801 cDNA sequence to the genome sequence revealed that the 3' end of the cDNA insert contained a segment that mapped to a different chromosome (i.e. Ch 3) than the CDS (i.e. Ch X) suggesting that it is a cloning artifact. The multiple encounters with artifactual sequences from EST clones during this project emphasize the necessity for verifying sequences prior to reaching any conclusions. In summary, *furrowed* is a paralogue of the CG9095 sharing all domain types but differing only in the number of CCP domains downstream of the lectin

domains. Both are genes are localized on the X chromosome (CG9095:13B1, *furrowed*: 11A2), suggesting that they arose through a duplication event. In the fruit fly, these are the only examples of proteins with this domain architecture and the only FBP-like domains found in the genome. The *furrowed* mutant was so named due to the crevices observed in the misdeveloped eyes of fruit flies with lesions to this gene. This phenotype along with aberrant bristles led to the conclusion that the *furrowed* gene was involved in ontogeny of sensory organs. More recent work detected strong expression of *furrowed* in primary pigment cells in the eye of adult fruit flies [248] demonstrating it is not restricted to developmental processes. Recently, peptidoglycan-binding receptors [249, 250] were identified in fruit fly as the missing pathogen receptor that activates the Toll pathway. It has also been discovered that the specificity of the immune response correlates with the source of the peptidoglycans [251]. Considering these developments, the homology of CG9095 with *furrowed*, and their similarity of expression suggests that CG9095 may rather play a role in development than in immunity. A unique feature of CG9095 is the presence of two different lectin domains with potential to bind Fuc. The description of fucosylated glycoproteins in *D. melanogaster* [252] confirms the presence of available ligands for these receptors. At this point in the studies, the binding site of FBPL was not known so it was not possible to predict lectin activity, but the residues of CTLDs that mediate binding have been identified [253]. It is possible to predict ligand-specificity and likelihood of binding [254] based just on consensus to these residues [83]. The result of this approach is that CTLDs have apparently diverged into those that bind sugar and those that retain the lectin fold but are unlikely to bind carbohydrate. An analysis of CTLDs present in *D. melanogaster* [84] indicated that the domain present in *furrowed* appears capable of binding carbohydrates while that of CG9095 does not. Closer inspection of the CG9095

CTLD indicates that it has several changes to the Ca^{2+} -binding residues that would preclude binding to carbohydrate (fig. 53). Specifically, there is a deletion of the segment containing the canonical Q-P-D and E-P-N motifs which determines specificity for galactose and mannose, respectively [255]. As documented for other CTLDs [256], it is possible that CG9095 will bind to proteins instead of carbohydrates.

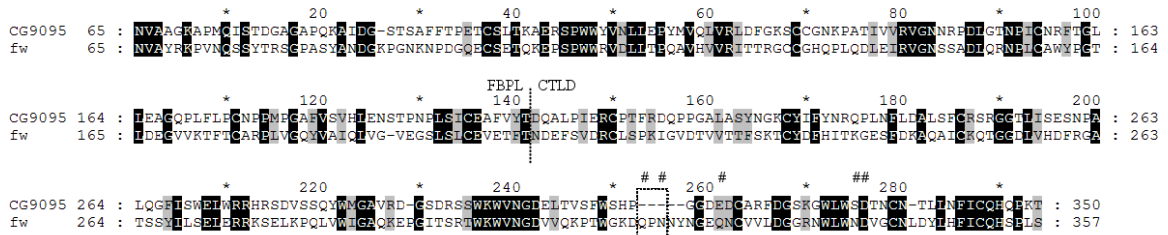


Figure 52. Alignment of the segments from CG9095 and *furrowed* spanning the FBPL and CTLD from the *Drosophila* homologues.

The dotted vertical line (|) marks the junction between the FBPLD and CTLD. The dotted box demarks the position of the Man/Gal-determining residues in the CTLD. The pound sign (#) indicates the CRD residues involved in binding Ca^{2+} . Black shading indicates identity. Gray shading indicates similarity. The segments share 31% identity and 50% similarity (Blosom 62).

The recent completion of the mosquito, *Anopheles gambiae*, genome [257] sequence allows its comparison to the fruit fly's genome [258]. A query of the mosquito genome indicates that it possesses both a CG9095-like gene and *furrowed*-like gene (GenBank Accession AAAB01008846.and AAAB01008811, respectively) syntenic on chromosome X, which supports an orthologous relationship. Despite an estimated 250 million years having passed since these groups diverged, the level of conservation between *furrowed* orthologues is higher (61% identity, 80% similarity) than the overall similarity reported (56%) between many fruit fly and the mosquito orthologues [258]. It would thus appear that these proteins are subject to stronger selective pressures. The genome of the honeybee (www.ncbi.nlm.nih.gov/BLAST/Genome/Insects.html) also encodes homologues of these receptors suggesting these receptors are conserved throughout dipterans and hymenopterans. Annotation of other insect

genomes [259] will help trace if this cell receptor is conserved throughout this class of arthropods or represents an innovation associated with the development of holometabolism. An FBPL from a chelicerate has already been described from the horseshoe crab, but is not orthologous to the insect receptor. Tachylectin-4, an FBPL purified and cloned from hemocytes of the Japanese horseshoe crab (*T. tridentatus*) [181], contains only a single domain, but forms larger noncovalent assemblies (470-kDa) than the bass and eel FBPLs. This lectin is one of many [63] characterized from secretory granules released from hemocytes upon stimulation with bacterial products. The appearance of horseshoe crabs in the fossil record dates back to the Cambrian period [260] and indicates they have changed little since. Therefore, in later appearing dipterans, the FBPL domain appears to have been co-opted to serve what is likely a cellular adhesion role.

The lack of any FBPL homologue in the genome of several nematodes screened [81, 261] contrasts with their presence in arthropods despite both being protostomes. This inconsistent distribution of FBPLs among protostomes may be better understood in light of revision of the relationship between arthropods and nematodes as ecdysozoans [262] rather than related as descendant and ancestor, respectively. This is part of a much broader reorganization of the metazoan evolutionary tree [263] which changes the perception of what were intermediary taxons into actually independently evolving branches. It is most likely that FBPLs were lost from the nematode lineage but remained in arthropods. The presence of FBPLs in other protostomes would support that the domain was present in the earliest ancestors of this group. Indeed, FBPLs are found in lophotrochozoans, the second major division of protostomes. ESTs from a planarian (GenBank Accession BP188149) and a bivalve mollusk (GenBank Accession CD647591 and CD646763) convincingly reveal an FBPL motif. In the

case of the oyster, the FBPL even appears to possess more than one domain. Consequently, from these results the phylogeny of FBPLs more likely appears to be a radiating bush rather than a progression of branches representing stages of domain evolution. Although the FBPL is perceivably ancient, it remains unknown if it was present prior to the emergence of Bilateria. The establishment of genomics projects in porifera and cnidarians (genome.wustl.edu/) should contribute to elucidate this question. Various genome projects in early branching eukaryotes such as plants, protozoa, and fungi [264] have not yielded any FBPLs. Apposite to this issue is a survey of protein domains in choanoflagellates [265], a group of unicellular and colonial flagellates which are considered an outgroup of animals. In contrast to the C-type lectin domain being present, no FBPLs are detectable in the substantial number of ESTs sequenced suggesting that this domain appeared in animals. The presence of the C-type domain in choanoflagellates is evidence of the early appearance of this domain and others prior to the development of multicellularity in metazoans yet after the appearance of opisthokonts (e.g. Fungi). Interestingly, one ORF presents a C-type domain associated with three CCP domains similar to the arrangement described in *D. melanogaster* CG9095. In summary, the sporadic distribution of FBPLs in protostomes parallels that in deuterostomes pointing to a preponderance of gene loss during this family's evolution.

Paucity of FBPs in prokaryotes: vertical vs. horizontal transfer?

Although the initial search of FBPLs only detected homologues in animals, the completion of *Streptococcus pneumoniae* genome [266] presented the first evidence of this domain's presence in prokaryotes. This important human pathogen has a special place in the history of molecular biology since it led to the discovery of DNA as the agent of heredity, facilitated by its ability for uptake of exogenous DNA [267]. Interest in the great diversity within prokaryotes and the

small size of genomes has resulted in the rapid completion of over 100 genomes [264]. Since the establishment of the Comprehensive Microbial Resource (CMR) [268], browsing the multitude of completed bacterial genomes is greatly facilitated. Using this resource a search of all ORFs was undertaken to test if any bacteria contained an FBPL domain. Surprisingly, an FBPL was detected only in the genome sequence (2.16 Mb) of the capsulated and virulent strain (TIGR4) of *S. pneumoniae*. The ORF, designated SP2159 (GenBank Accession AAK76213), consists of a 579 residue NH₂-terminal segment of no detectable homology followed by a triplet of FBPL domains. In contrast to the eukaryotes, the FBPL domains of SP2159 are devoid of Cys, which suggests that this protein is localized intracellularly where the negative redox potential abrogates any structural stabilizing effect of cystines, but the presence of a confidently predicted signal sequence cleavage site (VKA₃₁-D₃₂N) suggests otherwise. Indeed, FucU along with RbsD are the first intracellular Fuc-binding protein described in bacteria [269]. Interestingly FucU shares homology to genes in vertebrates suggesting they bind sugar as well. Since *S. pneumoniae* is a frequent cause of meningitis and pneumonia in humans, much attention is focused on searching for vaccine targets. One strategy relies on prediction of surface proteins by scanning ORFs for export signals [270]. SP2159 was not among 69 proteins predicted to be present on the surface of *S. pneumoniae* [266] so it remains to be determined exactly where this protein localizes. The location of SP2159, as is FucU, in the genome is within a regulon dedicated to Fuc catabolism [271]. *S. pneumoniae* is reported [272] to grow on diverse carbohydrates although not on fucose alone, yet analysis of the Fuc regulon promoter indicates that it is specifically activated by Fuc [273]. Therefore, it appears that the catabolism of Fuc solely cannot sustain growth due to requirement of an unknown product from glucose catabolism. Tettelin et al. [266] suggest that the fucose regulon may be involved

in the uptake of N-acetylfucosamine or other fucose derivatives for the purpose of capsule synthesis since N-acetylfucosamine has been isolated from the capsule of *S. pneumoniae*, but this appears improbable, as no Fuc salvage pathway such as found in mammals has been described in bacteria [274]. The sequencing of the avirulent, uncapsulated strain R6 (2.03 Mb) [275] allows complete genome comparison to the virulent TIGR4 strain. Indeed, the R6 strain also possesses an identical gene, spr1965 (GenBank Accession 15459666) suggesting this gene does not contribute to the differences in strain virulence.

Genome comparisons among prokaryotes facilitate domain annotation by revealing unexpected homologies. The release of the genome of the “flesh-eating” bacterium *Clostridium perfringens* (3.03 Mb) [276] revealed a putative protein, CPE0329 (trEBML Accession Q8XNK4), whose COOH-terminal contains a motif similar to the NH₂-terminal segment of SP2159 (fig. 54, Panel A). This shared domain stretches from Gln₃₆ to Ile₅₈₉ of the *S. pneumoniae* SP2159 protein, and shares 33% identical residues (fig. 54 Panel B). No other protein in the current gene databases exhibits this motif, and its function in *C. perfringens* is unknown. However, the presence of sugar catabolism genes surrounding CPE0329, like SP2159, suggests this novel domain also has a role in carbohydrate metabolism (www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl). In a different species of clostridia, a xylanase was described that contains triplicate family 6 carbohydrate-binding modules (CBM) [277] showing no sequence similarity to FBPLs. However, the avidity effect produced by the multiplicity of CBMs in the xylanase serves as a paradigm for the analysis of SP2159.

presented a single FBPL domain. It seems telling that the presence of FBPL domains in bacteria correlates with their broad ability to utilize diverse carbohydrates. The only other detected domain on the polypeptide is that of FA58C coagulation factor domain (GenBank Accession smart00231) which is upstream from the FBPL domain. Annotation of this domain states that it transferred from vertebrates to bacteria, but a more detailed examination reveals that the profile includes a fungal member, and therefore its presence in bacteria may truly reflect vertical transmission. In enzymes that target carbohydrates as substrates, such as glycooxidases and glycohydrolases, the apparent function of this domain [281, 282] is to bind and tether the carbohydrate substrate while the contiguous enzymatic domain exerts its particular activity. However, no detectable enzymatic domains are present in ZP_00065873.1 to support it being involved in degradation. Admittedly, the putative polypeptide is longer than just the two identified domains, leaving the possibility that an unrecognized enzymatic domain is present, or that a portion of the peptide even associates with a catalytic domain. The presence of cell wall protuberances observed on *M. degradans* suggest that it produces anchored cellulosome-like [283] complexes that facilitate the degradation of insoluble complex polysaccharides. Hence, it is possible that the FBPL of ZP_00065873.1 functions in binding cellulosomes to insoluble carbohydrate substrate, such as the terminal fucosylated xyloglucans [284] present in plants, to facilitate degradation. Remarkably, although the clostridial CPE0329 domain is not present in the *M. degradans* FBPL, *Clostridium* are also well known for forming cellulosomes [283].

Clearly, the presence of FBPLs in species within both Gram-positive, low G+C firmicutes (*S. pneumoniae*) and Gram-negative γ -proteobacteria (*M. degradans*), suggests that this domain either has undergone widespread loss or has been disparately transferred into bacteria in multiple

events. In contrast to the FA58C domain, the noticeable absence of FBPLs in the genomes from the many bacteria and fungi sequenced suggests it indeed may have horizontally transferred directionally from bilaterians to *S. pneumoniae* and *M. bulbifer*. The phenomenon of horizontal gene transfer (HGT) is generally accepted among Eubacteria and Archaeobacteria since it is supported by several examples from different gene families [285]. It is proposed that a lack of nucleus and their close association in communities produces the frequency of HGT amongst these cells. For the scenario of bacteria-to-eukaryote HGT, the maintenance of a bacterial endosymbiont has allowed subsequent movement of organellar genes (i.e. chloroplast and mitochondria) to the nucleus [286]. Recent controversy arose when bacterial genes were stated to have directly transferred to vertebrates [287] since this requires the improbable events of transfer to germ cells and eventual gene-fixation. This conclusion was promptly challenged with evidence of orthologues present in intermediate eukaryotes [288-290], in this case the slime mold. Examples supporting animal-to-bacteria HGT are few [291-293] and the mechanism by which it occurs is unknown but this transfer direction seems plausible especially for pathogens, like *S. pneumoniae*, in close association with their hosts. Presently, thorough phylogenetic analysis of SP2159's ancestry is not possible since no orthologues have been identified even among other Streptococcaceae [294]. Moreover, the genomes of the beta-hemolytic *S. pyogenes* [295], *S. agalactiae* [296], and *S. mutans* [297] do not present evidence of a Fuc regulon. The absence of FBPLs in these species suggests that SP2159 maybe solely present in the non-beta-hemolytic viridans *Streptococcus*, *S. mitis* group or even constrained to *S. pneumoniae* alone [298]. Since the *S. mitis* genome has been only partially sequenced, it is presently unknown if it shares the Fuc regulon of *S. pneumoniae*. The inconsistent presence of the Fuc regulon in bacteria suggests that gene loss for the pathway [299] is the cause for the

restricted species distribution of FBPLs. However, the identified Fuc regulons demonstrate a general distribution throughout many groups of bacteria (GenBank Accession COG4154 (FucU); COG0738 (FucP); COG2407 (FucI) and the FBPL is absent in all. Either this is the result of early gene loss prior to bacterial divergence, or the product of horizontal transfer as discussed above.

FBPL motif and domain topology

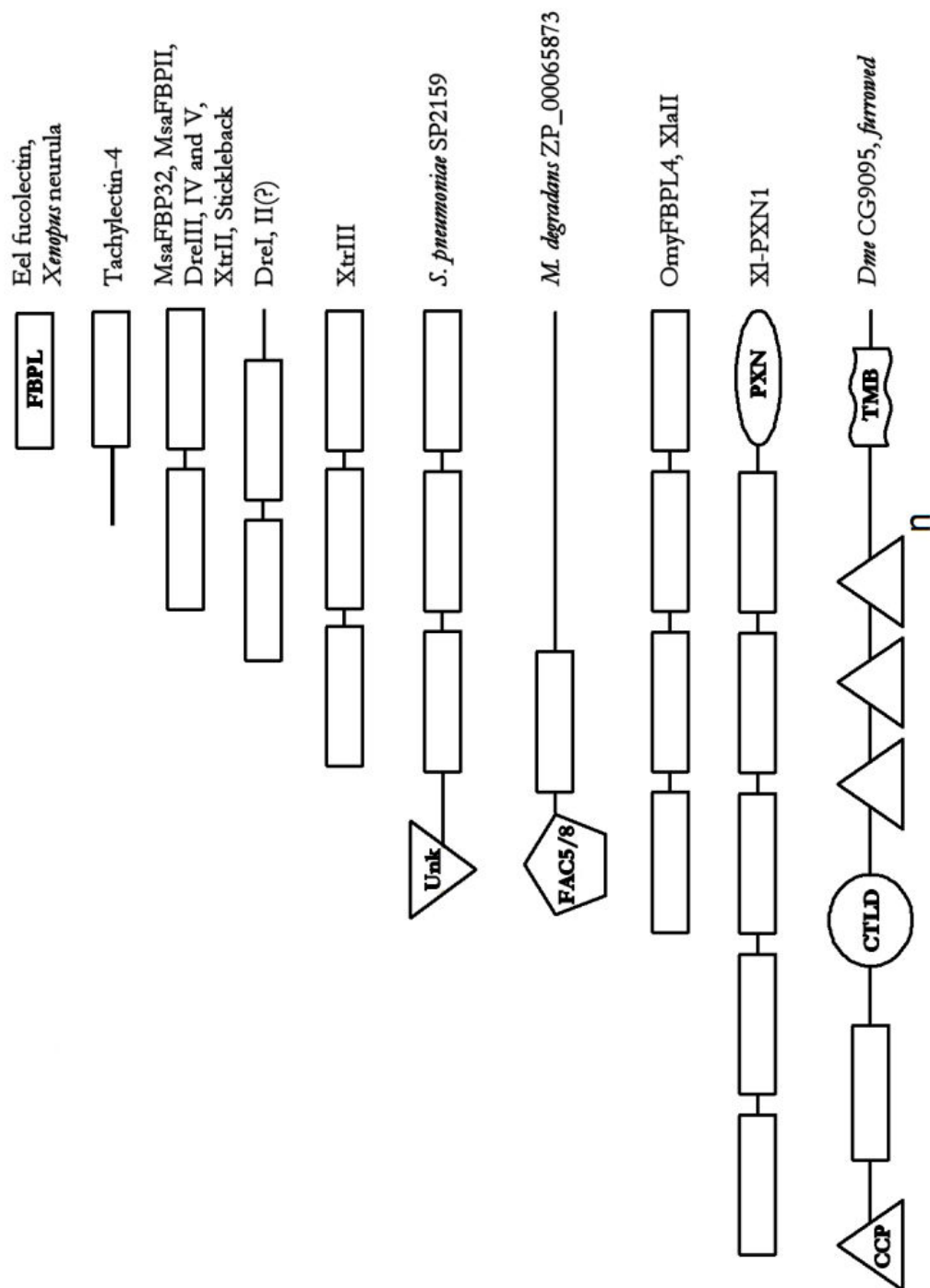
At the outset of this project, only one homologue (i.e. PXN1-XENLA) of the bass FBP32 was identified. Presently, forty FBPLs have been identified by sequence similarity from the major branches of the tree of life, including deuterostomes, protostomes and bacteria with exception of fungi and plants (Table 5). The motif shared by these proteins does not present similarity to any described lectin family, and therefore represents a novel lectin domain showing specificity for terminal Fuc. As mentioned earlier, the variable number of repeats of FBPLs and formation of mosaic combinations suggests that this motif represents an independently folding structural domain (fig. 54) [106]. Currently, domain concatenations range from singular up to five domains and possibly reflect the selection of varying levels of valence to control affinity. These tandem arrangements and multiple copies of FBPLs are evidence of the recurring role gene duplication [105] has played in shaping FBPL diversity. Several unique mosaic domain combinations [300] have been described from bacteria, fruit fly and frog, but none found so far in fish. Signal peptides on most FBPL ORFs predict that these lectins primarily localize as extracellular proteins. The structure, which the FBPL motif adopts, is unknown but it likely shares several common features observed in mosaic proteins [301]. Secondary structure within most domains of mosaic proteins consists of β -strands. It is frequently observed that

duplicated domains frequently have NH₂- and COOH-terminals entering and exiting, respectively, close in the structure [302] which would allow domains to extend from the polypeptide. Undoubtedly, the elucidation of an FBPL structure will settle many questions regarding this domain. It follows from this topological diversity that quaternary structure variation is bound to be high among these lectins. For example, the presence of a pentraxin domain suggests that Xla-PXN assembles into a toroid surrounded by FBPL domains.

Table 5. Taxonomic distribution and properties of deduced polypeptides containing an FBPL

Taxon [303, 304]	Species	Designation	FBPLs	Length (a.a.)	M.W. (Da)	pI	Site	Ref.
<u>Eubacteria</u>								
γ-Proteobacteria	<i>M. degradans</i>	ZP000065873	1	790	85737.3	4.72		
Lactobacillales	<i>S. pneumoniae</i>	Spn AAK76213	3	1007	114309.6	5.69		[266]
<u>Eukaryota</u>								
Platyhelminthes	<i>D. japonica</i>	Dja_BP188149	1	144	16615.6	6.30	head	
Mollusca	<i>C. virginica</i>	Cvi-gonad	3(?)	I	-	-	gonads	
Class Bivalvia								
Arthropoda	<i>T. tridentatus</i>	Tachylectin-4	1	232	26639.3	6.05	hemocyte	[181]
Class Merostomatata							granules	
Class Insecta	<i>D. melanogaster</i>	CG9095	1	925	100949.8	8.01	all life stages	
		<i>furrowed</i>	1	1174	128144.9	5.56	all life stages, primary pigment cells of adult eye	[247, 248]
	<i>A. gambiae</i>	Aga_ebiT8141	1					
		Aga_ebiP5322	1					
Echinodermata	<i>S. purpuratus</i>	Spu_larva	1 (?)	I	-	-	larva	
Chordata								
Class Osteichthyes								
Subclass Actinopterygii (ray-finned fish)								
Order Anguilliformes	<i>A. japonica</i>	AjaFTL-1	1	158	17034.1	5.54	liver	[182]
		AjaFTL-2	1	158	17256.3	5.94	liver	[182]
		AjaFTL-3	1	158	17294.3	5.84	liver	[182]
		AjaFTL-4	1	156	16948.9	5.94	gills	[182]
		AjaFTL-5	1	158	16958.9	6.06	gills	[182]
		AjaFTL-6	1	154	17135.1	6.20	gills	[182]
		AjaFTL-7	1	158	17336.5	6.06	gills	[182]
Order Cypriniformes	<i>D. rerio</i>	DreI-FBPL	2	345	37652.8	6.13	adult	
		DreII-FBPL	1(?)		I		adult	
		DreIII-FBPL	2	288	30970.8	4.65	gonads	
		DreIV-FBPL	2	293	32093.8	5.83		
		DreV-FBPL	2	290	31122.9	6.38	a. kidney	
	<i>C. carpio</i>	CaH170	1	139	15181.0	9.21	a. kidney	[231]
Order Salmoniformes	<i>O. mykiss</i>	OmyFBPL4	4	698	77288.0	5.19	liver	
Order Gasterosteiformes	<i>G. aculeatus</i>	Gac-CDA91-B12	2	292	31971.3	5.12	n/a	
		Gac-CDA35-C06	2	292	31793.2	5.83	n/a	
Order Perciformes	<i>M. saxatilis</i>	MsaFBP32	2	293	32393.1	6.16	liver	
		MsaFBPII	2	293	32304.1	5.74	liver	
	<i>M. chrysops</i>	MchFBP32	2	293	32308.0	6.01	liver	
		MchFBPII	2	293	32174.9	6.00	liver	
Order Tetraodontiformes	<i>F. rubripes</i>	Fru_1786	2	291	32158.6	5.21	heart, gonads	
		Fru_5138	2	285	31382.4	8.49	n/a	
	<i>T. nigroviridis</i>	Tni_794_4	2	287	31505.9	5.76	n/a	
		Tni_461_3	2	309	34285.4	6.08	n/a	
Subclass Sarcopterygii (lobe-finned fish)								
Class Amphibia								
Order Anura	<i>X. laevis</i>	Xla-PXN	5	1092	120426.2	5.14	liver	
		XlaII-FBPL	4	578	63255.0	5.29	liver	
		Xla-neurula	1	176	19336.9	7.74	embryo	
	<i>X. tropicalis</i>	XtrlI-FBPL	3	432	47365.3	5.98	embryo	
		XtrlII-FBPL	2	295	32764.5	5.21	embryo	

?: indicates that the cDNA is incomplete so the actual number of domains is not known



The increased representation of diverse FBPLs permits a comprehensive examination of the profile of conserved residues that define the motif (fig. 55). Among the most conserved residues are six Cys-containing positions present throughout the domains (i.e. MsaFBP32: Cys42, Cys74, Cys75, Cys97, Cys113, and Cys139). Based on this conservation pattern and results from SDS-PAGE, these Cys likely form intradomain disulfide bridges. Surprisingly, only binary FBPLs such as MsaFBP32 present a different combination of Cys conservation between domains. Although disulfide bridges have not been mapped in MsaFBP32, the differential presence of Cys pairs between the NH₂- and COOH-domains and the unique placement of contiguous Cys suggests likely bridging mates (i.e. Cys42-Cys139, Cys74-Cys75, and Cys97-Cys113). The presence of non-conserved Cys residues in odd numbers as in tachylectin-4, DreI-FBPL, DreII-FBPL, Aja-FTL-2, -3, and -7 suggests the existence of other bridges, some that may be between subunits. Several hydrophobic positions are also strongly conserved, underlining the critical structural role they likely play in folded protein. The unique PWW motif, so useful during the design of MOPAC primers, does show some variation, which would hinder the design of universal primers for FBPLs. Large gaps of variable width occur in areas showing low conservation in the alignment and likely correspond to loops, which are the most variable of secondary structures. Although several positions exhibit a high frequency of certain amino acids, it is impossible to predict from the consensus which positions interact with Fuc.

Prospects for tracing orthology between FBPLs are poor due to their inconsistent taxonomic distribution and differing topologies. However, the comparison of domains may reveal relationships related to their expansion. For this reason, an unrooted distance tree (fig. 56) was constructed from the alignment of all available FBPL domains. The resulting multifurcating

tree did not confidently resolve into deep nested clades but did rather indicate close relationships among several domains. One clade held all tandem domains from Xla-PXN suggesting they arose from internal domain duplications. A second clade groups all other *Xenopus* spp. FBPLs, reiterating independence from Xla-PXN and the teleost FBPLs. The orthology between the fruit fly and mosquito *furrowed* and CG9095 receptors supported by similarities in domain topology and chromosomal location is also sustained by their pairwise clustering. Curiously, the oyster FBPL groups with prokaryote FBPLs, which is most likely due to long branch attraction between the most divergent sequences. The exclusive clade containing the multiple copies of Japanese eel fucolectins strongly suggests their origins from duplications within the Anguilliformes lineage (i.e. eel). Greater similarity between FTL-2 and FTL-3 and between FTL-1 and FTL-5 suggests these are the most recent duplications in the group. The first pair is expressed in the liver like FTL-1 but FTL-5 was isolated from gills pointing to likely candidates for comparing fucolectin regulation. The presence of multiple FBPL copies contradicts earlier protein family analysis suggesting that duplication did not contribute significantly to gene expansion in eels [305]. One conspicuous feature of this tree is the two sister clades showing the grouping of N-terminal and COOH-terminal domains of binary FBPLs from acanthopterygians (stickleback, bass and pufferfish). However, the duplicate genes from each species tend to group together suggesting that independent duplication events have occurred along all three lineages. Applying character-based methods may improve the resolution of this gene tree but the extensive gene duplication and gene loss likely confounds any attempt at recovering reconstructing ancestry of the FBP domain.

Figure 55. Alignment of the FBPL domain.

Sequence names are composed of species name, protein name, if given, followed by the domain numbered from N to C terminus in the case of tandem. Invariant residues are shaded in black, conserved residues in ≥ 80 of sequences are shaded in gray with white lettering; conservatively substituted residues in $\geq 60\%$ of sequences are shaded in grey with black lettering. Sequence numbering starts from the first amino acid of the NH₂-sequenced protein or predicted cleavage site of the deduced full primary sequence. Abbreviations: Xla, *X. laevis*; Xtr, *X. tropicalis*; Msa, striped bass; Mch, white bass; Gac, stickleback; Omy, steelhead trout; Dme, fruit fly; Aga, tiger mosquito; Dre, zebrafish; Cca, common carp; Fru, tiger pufferfish; Tni, spotted green pufferfish; Aja, Japanese eel; Aan, European eel; Ttr, Asian horseshoe crab; Dja, planaria, Cvi, oyster, Mde, *M. degradans*; Spn, *S. pneumoniae*. Alignment was produced with Clustal_X v.1.81 [203] and shaded with GeneDoc v.2.6.002 [204]. Consensus is illustrated on bottom row by lowercase letters for the most frequent residue and numerals indicating Blossum 62 matrix similarity groups (i.e. 6: LIVM).

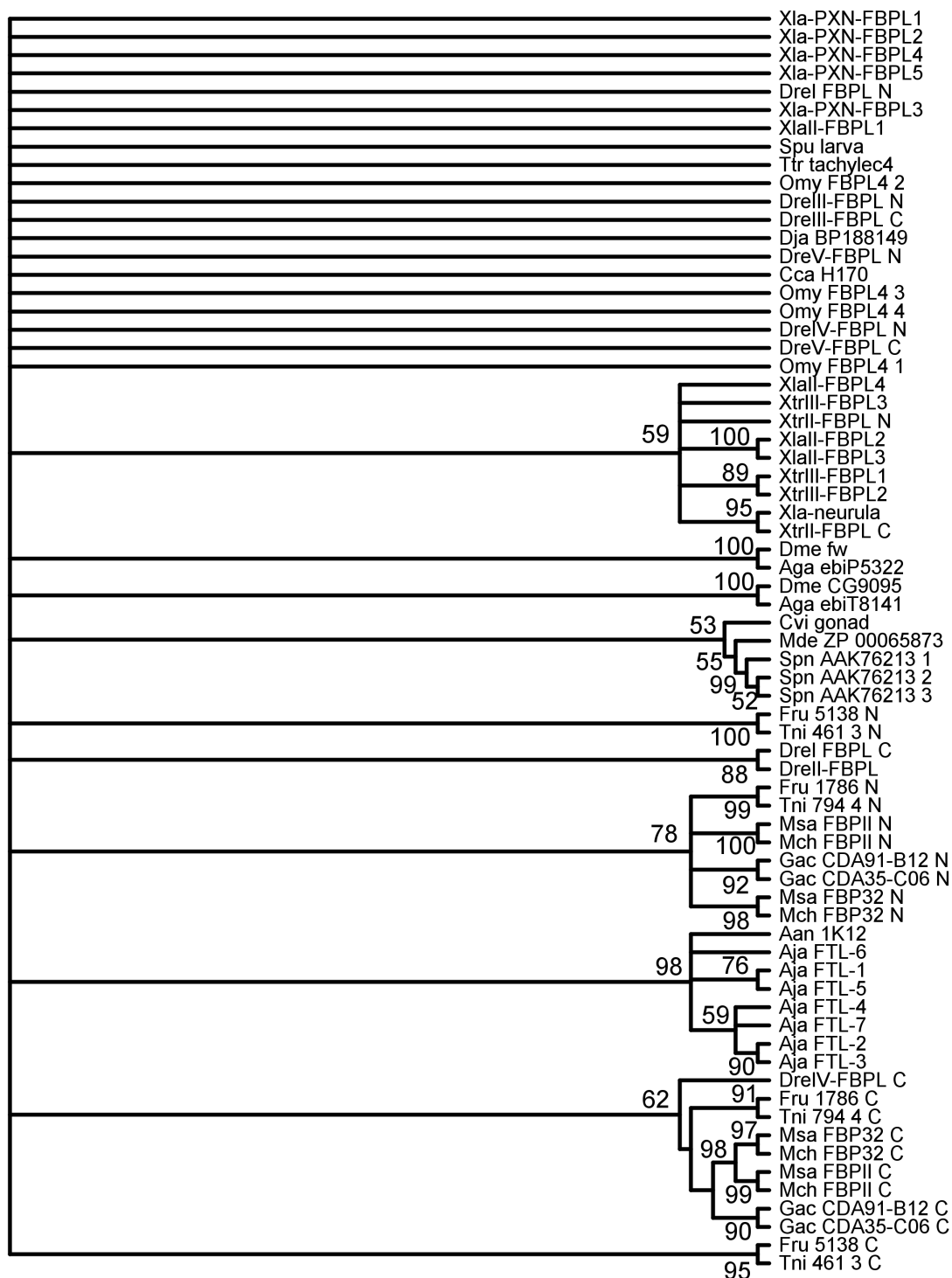


Figure 56. Analysis of genetic distance between FBPL domains. The phylogram was created from neighbor-joining analysis using Gonnet matrix as implemented in Clustal_X v.1.81. Distances were corrected for multiple substitutions and gap positions were excluded. Bootstrap values are percentages from 1000 iterations. Nodes with bootstrap values below 50% were collapsed.

Phylogenetic analysis of binary FBPLs

Sequence alignment of FBP domains indicated that many of the binary FBPLs discovered in teleosts were more closely related to each other, than to the rest of FBPLs. To analyze this relationship further, an alignment of all binary FBPLs was produced using the deduced full-peptide sequences that reflected the actual topological order of domains (fig. 57, Panel A).

Examination of the domains indicates that even between binary FBPLs there is much variation especially in the segments permissible to gaps. For example, the conserved pattern of putatively paired cysteines observed in the domains of bass FBPLs (i.e. Cys42-Cys139, Cys74-Cys75, and Cys97-Cys113) varies for the other homologues. The phylogram produced with these mature ORFs (fig. 57, Panel B) recovered the same clade seen for the COOH-terminal domains from bass, stickleback, pufferfish and one zebrafish FBPL. Again, the higher similarity between paralogues of bass and stickleback suggests that they emerged independently in each species lineage. The exception is the pufferfish, where greater similarity is observed between species indicating that duplication preceded their speciation. A similar topology was observed for the bass duplicates in the previous domain tree. The placement of DreIV, the gonadally expressed FBPL, as an outgroup to the bass, stickleback, and pufferfish clade suggests it is the ancestral form to the binary FBPLs present in the more derived fish. The interesting, albeit weak, association between DreI and XtrII suggests a possible link between the FBPLs of tetrapods and ray-finned fish. In summary, binary FBPLs have diversified through lineage-dependent gene duplications and speciation events producing a combination of paralogous relationships unique to teleosts.

A

[illegible]

B

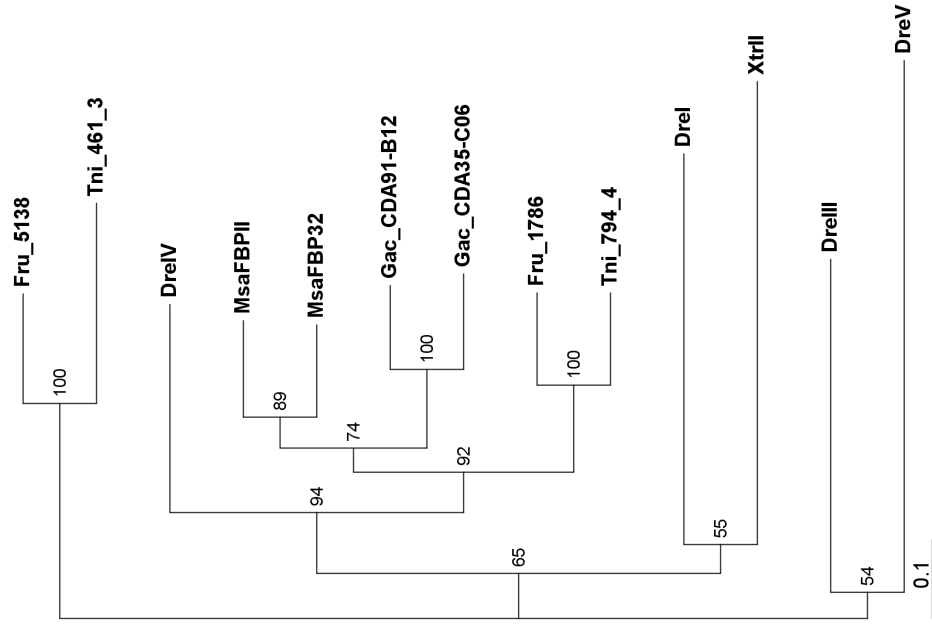


Figure 57. Analysis of evolutionary distance between binary FBPLs. Multiple alignment of full ORFs with identities and similarities shaded (see Fig. 55). Consensus line is formatted as follows: uppercase, 100% conservation; lowercase, majority residue; number, Blossum similarity group. **B.** The phylogram was created from neighbor-joining analysis using Gonnet matrix as implemented in Clustal_X v1.83. Distances were corrected for multiple substitutions and gap positions were excluded. Bootstrap values are percentages from 1000 iterations. Scale bar measures substitutions per site.

Unknown fate of FBPs during tetrapod evolution

One remarkable feature of the FBPL family is its apparent absence in higher homeotherm vertebrates. The presence of FBPLs in *Xenopus* spp. confirms that this protein family was firmly established into the tetrapod lineage, but it is presently unclear up to what stage of tetrapod evolution it persisted. Currently, no FBPLs are detectable by sequence alignment in the finished human genome (www.ncbi.nlm.nih.gov/genome/guide/human/) or the draft mouse genome [306] and annotated cDNAs [307] (www.ncbi.nlm.nih.gov/genome/guide/mouse/) raising the possibility that FBPLs have been lost in higher vertebrates. For humans, more than 99% of the euchromatin (i.e. gene containing DNA) has been sequenced and assembled into an accurate finished sequence, so if any FBPLs were present they most likely would appear during a sequence similarity search.

One of the factors correlated with increased organismal complexity is the increased diversity of shuffled protein domains [308], so it is surprising to find that a protein family already established and highly diversified in the tetrapod lineage appears to have disappeared from higher vertebrates. Comparative genomic studies suggest that gene loss after duplication [309] is a frequent event even in vertebrates [310] but the loss of a protein family has not been documented. The absence of FBPLs in these groups may not represent the situation for all mammalian taxa since recent evidence points to primates and rodents being more closely related than to other mammals (i.e. carnivore and ungulates) [311]. The rising availability of mammalian genomic resources [312], encouraged by the increased interest in comparative studies, eventually will permit addressing if other eutherians, marsupials and monotremes also

lack FBPLs. Reptiles, as predecessors of mammals, are of interest for determining if FBPs persisted after the emergence of tetrapods on land.

Unfortunately, reptiles as a group are currently unrepresented in genome studies (igweb.integratedgenomics.com/GOLD/) so the presence of FBPLs in this order is cannot be readily addressed. Revisions in amniote phylogeny based on morphological and molecular characters places birds as a sister group of reptiles, although their affinity among reptile groups is still debated [313]. A query of an extensive chicken EST collection from various tissues [314] did not return any matches so it appears that FBPLs are also absent from birds and likely reptiles. The production of publicly available BAC libraries from representatives of extant reptilian clades (i.e. a Testudine, Crodilian, Squamate and a Rhynchocephalian) is currently planned (www.nsf.gov/bio/pubs/awards/bachome.htm), which may enable a broader FBPL survey in reptiles. The lack of results tentatively rules out the presence of FBPLs in land-dwelling tetrapods, but the paucity of the FBPLs observed in invertebrates weakens this conclusion. One explanation for not detecting FBPLs in the above mentioned taxa with the approaches selected, may be that they have been subject to rapid diversification in these groups, and have diverged beyond statistically significant recognition [315]. However, despite the extensive time (>500 Ma) passed since the protostome/deuterostome divergence [316], the FBPLs of ectothermic vertebrates and arthropods still share significant similarity. Examining the ubiquity of FBPLs in extant amphibians, including “modern frogs” (Neobatrachia), and methodically examining their presence in reptiles (anapsids and diapsids) may help in addressing this apparent FBPL family extinction, but the unclear evolutionary path from amphibians to reptiles [317] may still confound any conclusion for higher vertebrates. As for mammals, representation of vertebrate diversity in genomic databases will eventually grow

allowing a more thorough search for members of this lectin family. There is the possibility that apparent lack of FBPLs is correlated with the appearance of homeothermy since immunocompetence of cold-blooded vertebrates has been related to body temperature [318]. However, maybe a more dramatic evolutionary innovation that predates homeothermy, the development of the cleidoic egg, may be the demarcation between the vertebrate classes possessing or lacking FBPLs. This characteristic is believed to have allowed cold-blooded amniotes to leave behind a water-dwelling lifestyle and colonize land since their eggs no longer required water for developing. Again, a thorough survey of reptilian and amphibian orders for FBPLs is required to address this apparent lectin family extinction.

The lack of correspondence among available vertebrate FBPLs confounds the identification of the ancestor gene that gave rise to this diversity. Completion of an ascidian (urochordate) genome [319] and the advancement of an echinoderm genome [320] provided an opportunity for analyzing the complement of FBPLs in taxa representing early stages in the evolutionary path towards the emergence of vertebrates. Both of these organisms are invertebrates, but like vertebrates, they are deuterostomes. Specifically, echinoderms as well as hemichordates (e.g. acorn worm) represent the most primitive of extant deuterostomes, the non-chordates.

Likewise, the urochordates are the most primitive of extant chordates [321]. A search of the ascidian genome or EST collections [322, 323] did not produce any significant matches to FBPLs. Curiously, a plasmid clone (fig. 58, Panel A) produced for the genome-sequencing project (ghost.zool.kyoto-u.ac.jp/indexr1.html) exhibited a partial FBPL motif. The strong similarity of this putative ORF to FBPs suggests that it is authentic (fig. 58, Panel B) but the absence of this sequence from the assembled genome raises doubts. Annotation of this clone is woefully lacking but it apparently was among the Japanese-produced libraries that were not

included during the assembly of the *C. intestinalis* genome [319] (as described in the supplementary information). Evidently, further work is required to confirm if this clone is truly part of the ascidian's genome, so at present it appears that an FBPL is absent in the ascidian lineage. This apparent complete loss of the family is intriguing in light of the expansion of immune-related genes detected in *C. intestinalis*, including genes encoding surface receptors with multiple C-type lectin motifs [85]. Ascidians appear to be particular among urochordates, which may help explain the loss of this family in *C. intestinalis*. The notochord evident in the motile larval stage of ascidians unarguably places them morphologically among chordates, however upon morphing into an adult this structure disappears, as the individual becomes a sessile benthic inhabitant. Larvaceans, suggested to be the earliest diverging urochordates [324], are pelagic and retain their notochord into adulthood; therefore, the primitive adult body plan of ascidians appears to be secondarily derived. Despite being phylogenetically distant, one a deuterostome the other a protostome, *C. intestinalis* shares with *C. elegans* several embryonic characteristics such as rapid development, early fixation of cell fate and reduced embryo cell number [325]. Interestingly, these morphological similarities are paralleled by changes in their genomes. In contrast to other metazoan genomes analyzed to date, both animals possess a reduced nuclear-genome size manifesting high rates of mutation including gene order rearrangements and apparent gene loss (e.g. Hox genes). Both these physiological and genome characteristics suggest that both taxa have regressed relative to their respective ancestors.

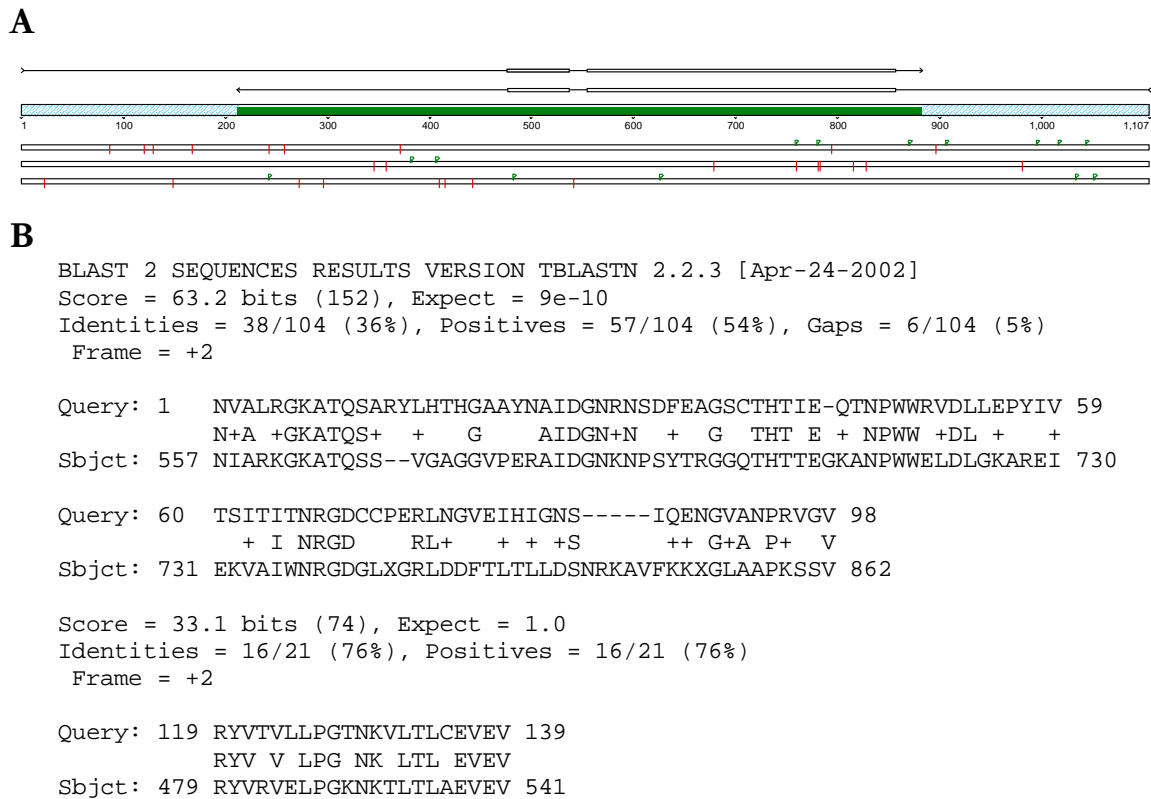


Figure 58. Tentative identification of an FBPL sequence in the ascidian, *C. intestinalis*.

A. Overlap of sequencing traces from both ends of the genomic DNA plasmid clone, GCiWno578_c20. Boxes indicate the FBPL motif presented in the lower panel. Note that the motif corresponding to the COOH-terminal end of the FBPL domain is upstream of its amino terminal motif. **B.** Blast2 pairwise alignment between the NH₂-terminal FBPL of MsaFBP32 and the translated sequence of genomic clone in panel A. In the intervening row, identical residues are indicated by letters and similarities by crosses.

Although FBPLs appear to be absent from ascidians, there is evidence of their presence in an

earlier diverging deuterostome, an echinoderm. Specifically, the purple sea urchin

(*Strongylocentrotus purpuratus*) possesses multiple FBPL ESTs expressed in larvae (GenBank

Accession CD310754, CD289590, CD307188, and CD304127). Similar to ascidians,

echinoderms appear to have regressed to an ancestral state, in this case that of radial symmetry;

nevertheless, they have retained the FBPL family. The spotty distribution of FBPLs in

deuterostomes resembles the scenario observed in protostomes suggesting a relaxation in

selection that eventually leads to gene loss. The expression of FBPLs during early sea urchin

development, similarly as in *Xenopus* spp., suggests their ancestral role is in deuterostome

development. However, an argument against this hypothesis is that genes with function(s) in development typically are phylogenetically widespread due to their indispensable role. In summary, an FBPL is absent from the assembled ascidian genome, but the tentative identification of a partial motif in an ascidian genomic DNA clone raises the possibility that the assembly is incomplete and that FBPLs indeed have not been lost from ascidians. Further work is necessary to confirm if this is so. Clearly, evidence from an echinoderm demonstrates the conservation of FBPLs in an early deuterostome, however completion of a cephalochordate's genome, the sister clade of vertebrates [321], may provide a better representation of the FBPL gene complement just prior to the emergence of vertebrates.

Displacement of FBPLs by C-type lectins?

The shared ligand specificity of FBPLs with C-type lectins raises the possibility that intragenomic competition [302] may be responsible for disappearance of FBPLs. In the extensive nematode genomes sampled, FBPLs are noticeably absent, but CTLDs have greatly diversified [83]. Indeed, CTLDs have proliferated more in nematodes than in the fruit fly [84], which still have an FBPL homologue. This concept of domains competing arises from the observation that some proteins (e.g. protein kinases and PH domain family) in time come to dominate a particular function despite the initial availability of functionally similar but unrelated domains. The hyperadaptability of these domains likely leads to establishing their dominance. Comparison of CTLDs identified in genome sequences available hitherto indicates an independent expansion and diversification of CTLDs [81-83, 254, 326] within each bilaterian lineage. Many of these domains are unlikely to bind carbohydrates since they are missing the canonical sugar-binding residues. The most dramatic example is illustrated by the antifreeze proteins from arctic fish, such as the herring and sea raven, where CTLDs no

longer bind sugars but rather bind to and disrupt the growth of nascent ice crystals [327]. The ability of CTLDs to evolve diverse specificities beyond carbohydrates may have provided the advantages to replace functions performed by FBPLs in the mammalian and nematode genomes. Alternate evidence for this non-orthologous gene displacement [328] is present in bacteria where eubacterial enzymes have been lost, but whose function has been replaced by unrelated archaeal functional analogs.

CHAPTER 4. GENETIC ANALYSIS OF FBPL FAMILY

The previous chapter demonstrated that MsaFBP32 is but a single example of a diverse and taxonomically broadly distributed newly-identified lectin family.. This was facilitated by the increased interest in sampling active transcriptomes in lieu of sequencing lengthy eukaryotic genomes. However, one disadvantage of focusing on cDNAs is that this approach fails to provide important information such as gene organization and context, which remain unknown since this information is lost upon transcription. The earlier failure of confidently reconstructing the ancestry of FBPLs suggests that qualitative characters such as shared genetic context may be helpful in assigning orthologies within the family. However, presently no studies have described the structure of the gene encoding any of the members of this lectin family. Since none of the sequenced genomes available at the time presented evidence of an FBPL gene, no observations could be readily made about them. This chapter presents the completion of the MsaFBP32 gene sequence and discovery of its nearby paralogue. In addition, it describes the diversity of FBPLs present in the finished genome project from pufferfish and the ongoing project for the zebrafish.

Materials and Methods

Screening of striped bass genomic DNA bacteriophage library

A striped bass genomic DNA bacteriophage library, kindly provided by Dr. Yonathan Zohar (COMB), in 20 separate pools after a single amplification was screened by conventional DNA hybridizations methods. The vector used for making the library was λ FixII (Stratagene, La Jolla, CA, USA), and the strain used for initial transfection was XL1-Blue MRA (P2)

(Stratagene). In preparation for plating plaques, the bacterial host, XL1-Blue MRA (Stratagene), was grown in Luria-Bertani broth/0.2% (w/v) maltose/10 mM MgSO₄ to an OD₆₀₀ of 1 at 37 °C shaking at 300 rpm, to an OD₆₀₀ of 1. The cells were pelleted at 1,000 x *g* for 5 min and suspended in sterile 10 mM MgSO₄ to an OD₆₀₀=0.5. Of this cell suspension, 600 µl was added to 12 ml polypropylene culture tubes (Fisher Scientific), and approximately 40,000 PFU were added to each tube and mixed. The transfection was incubated at 37 °C in a water bath for 20 min. Seven ml of Top agarose (Luria-Bertani broth/0.4% (w/v) agarose) equilibrated at 49 °C were added to each tube, and the contents were plated promptly by swirling on 150 mm Petri plates containing Luria-Bertani agar equilibrated at 37 °C. The plates were set on the bench for 15 min, and then transferred to a 37 °C incubator for lawn growth. After approximately 9 h the plaques had reached a diameter of approximately 1 mm and the plates were removed from the incubator and chilled at 4 °C for 30 min to harden the agarose top layer in preparation for plaque lifting. Positively charged nylon HyBond N+ membrane discs (132 mm) (Amersham Biosciences) were placed onto each plate and left for 1 min to adsorb the plaque DNA. An 18-gauge hypodermic needle was used to poke orientation holes through each membrane into the agar below for later orientation of the autoradiograph to the plates. The membranes were placed on sheets of Whatman 3MM filter paper (Fisher Scientific) wetted with 1.5 M NaCl, 0.5 M NaOH for 30 sec for bacteriophage lysis and denaturation of the double strand DNA, and then transferred to filter paper wetted with 0.5 M Tris-HCl (pH 7.4), 1.5 M NaCl for neutralization. Subsequently, the membranes were transferred to a tray of 2X SSC (20X SSC: 3M NaCl, 0.3 M sodium citrate) to clear them of any debris from the plates. The plaque lifts were dried and stored at room temperature between filter paper until hybridization.

Probe hybridization was performed using conventional methods as follows. The 480 bp probe used to screen the library was derived from a cDNA cloned and spanned the second tandem domain of MsFBP32. The probe template was generated by PCR amplification from a striped bass cDNA clone (pFBP6) clone with primers FBP469.F (5'-GTT ACG TGA CTG TGC TTC TAC CTG G-3') and FBP949.R (5'-ACA GGG TCA GGT ACT CTT CTC TTC C-3'), followed by gel purification of the amplicon with Qiagen's QIAquick Gel Extraction kit (Valencia, CA, USA). The PCR reaction mix was composed of 1X EnzOne 2000 buffer, 1.5 mM MgCl₂, 0.2 mM dNTP, 0.5 mM of each primer, 10 ng pFBP6, 2.5 U EnzOne 2000 polymerase (ID Labs), in a 100 µl final reaction volume within 500 µl PCR tube (Marsh). The cycling parameters were: (1) 94 °C , 2 min; (2) 94 °C , 10 sec; (3) 50 °C , 30 sec; (4) 70 °C, 1 min (repeat step 2 through 4 for 29 cycles); (5) 72 °C for 1 min. The DNA was radioactively labeled by random hexamer primer as described in chapter 2. The membranes were prepared by first wetting in 2X SSC and prehybridized in 80 ml of hybridization solution (hyb solution) in a round plastic storage container. The labeled probe was added to the solution at a concentration of 1x10⁶ CPM/ml and hybridization proceeded at 65 °C with rocking overnight in a forced air oven. Probed membranes were washed twice in 200 ml of 2X SSC, 0.1% (w/v) SDS for 15 min at room temperature. For increased stringency, the membranes were washed twice in 200 ml of 0.1X SSC, 0.1% (w/v) SDS at 65 °C for 15 minutes. Washed membranes were blotted to remove excess liquid and wrapped in Saran wrap for exposure to XAR10 film (Eastman Kodak, Rochester, NY, USA) overnight at -80 °C with an intensifying screen. The film was developed automatically in an XOMAT machine (Eastman Kodak, Rochester, NY, USA).

Phages that yielded a positive signal by autoradiography during the first screen were rescreened twice more by plating well separated plaques (approx. 100) on 100 mm LB Petri plates, and lifting to 82 mm Hybond N+ (Amersham Biosciences) membrane discs. The membranes were treated as described above for bacteriophage lysis and hybridized to the same probe. Phages that hybridized positively throughout the three purification screens were analyzed by restriction enzyme fingerprinting. Bacteriophage DNA was purified using Qiagen's Lambda Midi Phage Kit, according to the manufacturer's instructions. Liquid lysates (50 ml) were prepared in NCZYM broth by combining phage and bacteria (1:200), and incubating at 37 °C for 8 h. Approximately 5 µg of purified DNA was digested with 10 U of restriction enzyme at 37 °C for 1 h. The digested DNA was resolved on a 1% agarose, 1X TAE gel. A representative bacteriophage was picked for each restriction pattern identified as unique. Each of these clones was sequenced from both ends of the clone using T7 and T3 phage-polymerase primers to test if they were contiguous.

Southern blot analysis

The digested bacteriophage DNA resolved by agarose slab electrophoresis performed above was transferred to BioRad ZetaProbe GT membrane. To fragment large DNA bands, the gel was soaked in 0.25 M HCl for 10 min. Upward alkaline transfer using 0.4 M NaOH was performed according to the manufacturer's instructions. After overnight transfer, the membrane was washed in 2X SSC, and air-dried.

PCR amplification of MsaFBP32 gene from λMs15

A segment of the λMs15 bacteriophage clone was amplified by long-PCR. In a final reaction volume of 25 µl, using buffer #3 from the Expand Long PCR polymerase mix (Roche

Molecular Biochemicals; Indianapolis, IN, USA), the largest amplicon obtained used FBP1.F (5'-CTG GAC TCC AGG GAT AAA AGA TCT G-3') and HiT3 (5'-AAT TAA CCC TCA CTA AAG GG-3') primers by the following cycling parameters: (1) 94 °C , 2 min; (2) 92 °C , 10 sec; (3) 50 °C , 30 sec; (4) 70 °C, 12 min (repeat step 2 through 4 for 29 cycles); (5) 70 °C for 12 min. The amplicon was subcloned into pCR2.1 vector using the Invitrogen's TOPO TA and transformed into *E. coli* TOP10F' host supplied with the kit.

Deletional directional sequencing

A CsCl-purified plasmid preparation [158] was used for preparing directional deletion constructs. The procedure made use of the Fermentas ExoIII/S1 Deletion kit (Hanover, MD, USA). Twenty µg of plasmid were digested overnight with 10 U of *Sac*I (New England Biolabs) in 50 µl at 37 °C. To produce a 5' overhang sensitive to the nuclease cocktail, a sequential digest with 20 U *Spe*I was performed for 2 hrs at 37 °C. The double-digested plasmid was purified by phenol/chloroform extraction and ethanol-precipitated in preparation of the nuclease treatment. Directional deletion by nuclease digestion was done according to the kit manufacturer's instructions.

Bacteriophage clone shotgun sequencing

To sequence the bacteriophage clone, the insert was randomly fragmented using partial digestion with *Sau*3AI, and subcloned by blunt end ligation into pBluescriptII plasmid [156]. Random clones were then picked for sequencing until the complete sequence of the phage insert was obtained. In detail, 10 µg of λMs15 clone DNA were digested with *Nco*I to release the bacteriophage vector arms. The insert DNA was gel purified on 0.7% low-melt agarose/1X TAE slab electrophoresed at 3 V/cm. The insert DNA-containing band was cut

into small pieces, weighed and digested with Agarase (New England Biolabs) at 42 °C for 1 hour. The released DNA was ethanol-precipitated and suspended in TE (pH 8). Purified DNA (5 µg) was diluted to 100 µl in 1X restriction buffer provided for the enzyme in preparation for random fragmentation. Optimal fragments for cloning into a plasmid vector (i.e. 2-9 kb) were obtained by titrating the restriction enzyme, *Sau3AI* (New England Biolabs), with equal amounts of insert DNA. Specifically, 30 µl aliquots of DNA solution were dispensed into the first of five 1.5 ml microtubes. The next three tubes held 20 µl and the fifth tube contained 10 µl of insert DNA. All tubes were chilled on ice prior to adding the restriction enzyme. Four units of restriction enzyme were added to the first tube followed by serially transferring 10 µl to the following tube, so that all tubes held 20 µl final volumes. Reactions were then incubated at 37 °C for 15 min on a heat block. The digestions were stopped by placing tubes on ice and adding 2 µl of 0.5 M EDTA (pH 8.0), followed by heat-killing the enzyme at 70 °C for 15 min. After analysis of 1 µl aliquots from each digest on 0.7% agarose/1X TAE those that presented the desired size range smear were pooled and phenol/chloroform-extracted, followed by ethanol precipitation.

The ends of the digested bacteriophage insert were repaired using DNA polymerase and polynucleotide kinase. The precipitated digested DNA was resuspended in 22 µl of TE (pH 8.0). The following was added in order: 3 µl 50 mM MgCl₂, 3 µl 0.5 mM dNTP mixture and 2 µl bacteriophage T4 DNA polymerase (5 U/µl) (New England Biolabs). The reaction was incubated for 15 min at room temperature. One µl of Klenow fragment *E. coli* polymerase I (5 U/µl) (New England Biolabs) was added followed by 15 min more at room temperature. The

volume was raised to a final 50 μl with water and phenol/chloroform extracted and ethanol precipitated. The repaired fragment ends were phosphorylated in preparation for vector ligation by resuspending the DNA in 23 μl of water, and adding in order the following: 3 μl 10X bacteriophage T4 polynucleotide kinase buffer, 20 mM ATP, and 1 μl bacteriophage T4 polynucleotide kinase (1 U/ μl). The reaction was incubated at 37 °C for 30 min followed by heat-killing of the enzyme at 65 °C for 5 min. The DNA was cleaned by phenol/chloroform extraction and ethanol-precipitated. The end-repaired/phosphorylated DNA was resuspended in 25 μl of TE (pH 8.0) in preparation for the insert-vector ligation step.

In preparation for insert ligation, pBluescriptII SK+ (Stratagene) was digested with a restriction enzyme that produces blunt-ends receptive to the blunted DNA fragments. Ten μg of plasmid were digested with 20 U of *HincII* (New England Biolabs) in a final reaction volume of 100 μl at 37 °C for 2 hours. The fully digested plasmid was dephosphorylated to avoid religation by adding 10 μl of water to the digested plasmid (100 μl), 10 μl of 10X CIP buffer and 5 U of calf intestinal phosphatase (CIP) (1 U/ μl) (New England Biolabs). The reaction was left for 1 hr at 37 °C after which the enzyme was removed by phenol/chloroform extraction and ethanol precipitation. Plasmid DNA was resuspended in TE (pH 8.0) at a final concentration of 100 ng/ μl .

For producing the shotgun library, ligation of the digested DNA to the plasmid vector was performed as follows. The reaction consisted of the following reagents: 5 μl of 100 ng/ μl plasmid, 0.167 pmole repaired DNA (approx. 0.1 pmole/ μl), 5X blunt ligation buffer (200 mM Tris-HCl (pH 7.5), 50 mM MgCl_2 , 50 mM DTT, 10 mM ATP, and 30% (w/v) PEG8000),

0.3 μ l bacteriophage T4 DNA ligase (400 U/ μ l) (New England Biolabs) and water to a final volume of 20 μ l. Incubation was overnight at room temperature. Transformation into chemically competent *E. coli* TOP10F³ cells (Invitrogen) used 2 μ l of the ligation reaction and 50 μ l cell aliquots. The steps in transforming the cells after adding 250 μ l of SOC broth (2% (w/v) tryptone, 0.5% (w/v) yeast extract, 0.05% (w/v) NaCl, 20 mM glucose) were 30 min on ice, 30 sec at 42 °C in water bath, 2 min on ice and 1 hr of recovery at 37 °C. Cells were finally β -gal selected by plating on LB agar/100 μ g/ml ampicillin/ 0.025 % (w/v) X-gal/ 0.05 % (w/v) IPTG plates. White colonies were picked and plasmids extracted by Wizard SV plasmid mini-preps (Promega) were sequenced using vector-based primer T7 (5'-TAA TAC GAC TCA CTA TAG GG-3') and T3 (20-mer) (5'-AAT TAA CCC TCA CTA AAG GG-3').

Cloning of MsaFBP II cDNA from liver

MsaFBP II was PCR-amplified from *M. saxatilis* liver cDNA with the following primers: BreIIA.F (5'-CGC TAA AGA GAA ACG CCT CT-3') and BreIIC.R (5'-TCG GTC TGC GAC GCT GTT GTT AA-3'). Transcript-s 3' and 5' ends of MsaFBP II were obtained by RACE with the following primers: MsfbpII3UTR.F (5'-CAG ACA ATA CCG CAT ATG TGC T-3'), MsfbpIIB.R (5'-TGA GCA CAT ATG CGG TAT TGT CTG-3'), and MsfbpIIA.R (5'-CAC ACG TGA TTG TTG CAC CTG AGA C-3').

PCR amplification of Mcfbp32

To amplify the orthologous white bass gene, DNA isolated from red blood cells was used as template. The primers used were FBP1.F forward primer (5'-CTG GAC TCC AGG GAT AAA AGA TCT G-3') and FBP949.R reverse primer (5'-ACA GGG TCA GGT ACT CTT CTC TTC C-3'). The PCR reaction mix consisted of 1X kit buffer 3, 0.2 mM dNTPs, 0.3 μ M

each primer, 0.5 µg *M. chrysops* DNA, 0.75 µl Expand Long polymerase mix (Roche Molecular Biochemicals) in a 50 µl volume. Cycling parameters were: (1) 94 °C, 2 min; (2) 94 °C, 10 sec; (3) 60 °C, 30 sec; (4) 70 °C, 5 min (repeat step 1 through 3 for 29 cycles) (5) 70 °C for 7 min. The resulting amplicon was cloned directly from the PCR reaction using TOPO TA kit (Invitrogen). The resulting clone (pMc.3) was sequenced by ExoIII/S1 directional deletion generated subclones as described below.

Isolation of genomic DNA

Fresh whole blood from white bass was collected using EDTA (pH 8.0) as anticoagulant (final concentration 10 mM) from caudal vein puncture of the white bass. DNA was isolated from the nucleated red blood cells employing the GenomicPrep Blood DNA Isolation kit (Amersham Biosciences) based on high salt precipitation of protein.

PCR amplification of FBP32 gene in hybrid bass

Striped bass, hybrid bass, and white bass genomic DNA was isolated from blood as described above. The striped bass intron-specific upstream primer was fbpI312.F (5'-GCA GTA TTA TTC AAC CCC CAG CTT C -3'). The white bass intron-specific upstream primer was fbpII1464.F (5'-GAC AAA CCA TGC AGA AGT TAT TAG CC-3'). Both primers were used in combination with the downstream primer FBP949.R (5'-ACA GGG TCA GGT ACT CTT CTC TTC C-3)', which is not species-specific. The PCR reaction mixture was 1X DyNAzyme EXT Optimized buffer, 0.36 mM dNTPs, 2 mM MgCl₂, 1 µM each primer, 2 µg of genomic DNA, 1U of Finnzyme DyNAzyme EXT polymerase mix (MJ Research, Waltham, MA, USA). The cycling parameters were: (1) 94 °C, 10 sec; (2) 94 °C, 10 sec; (3) 60 °C, 30 sec; (4) 70 °C, 10 min (repeat step 1 through 3 for 29 cycles) (5) 70 °C for 7 min.

Screening of zebrafish genomic DNA bacteriophage library

A λ FIX II bacteriophage genomic library from the zebrafish, *D. rerio*, was purchased from Stratagene. The probe used for screening the library was produced by PCR amplification from the pBre3' RACE clone with the primers, Bre2.F (5'-CAT CAC TAA CAG ACT GGA CTG CTG C-3') and MCSRACE. The cycling parameters were: (1) 94 °C, 2 min; (2) 94 °C, 15 sec; (3) 60 °C, 15 sec; (4) 72 °C, 1 min (repeat step 1 through 3 for 29 cycles) (5) 72 °C for 5 min. The PCR reaction mix was composed of 1X EnzOne 2000 buffer, 1.5 mM MgCl₂, 0.2 mM dNTP, 0.5 mM of each primer, 10 ng pFBP6, 2.5 U EnzOne 2000 polymerase (ID Labs) in 100 μ l. The library was screened following the procedures described for the screening of the striped bass genomic library. After Southern analysis of the positive clones using the same probe used in screening the library, labeled bands were identified for subcloning. For purification of bands for subcloning, the clone insert (10 μ g) was released with 20 U of *Xba*I or *Xho*I (New England Biolabs) at 37 °C for 2 h in a 200 μ l reaction volume. The digested DNA was purified by phenol/chloroform extraction and ethanol precipitated. To purify the desired bands, the digested DNA was separated on a 1% (w/v) agarose, 1X TAE slab gel and resolved bands were purified with Qiagen's QIAquick gel extraction kit. Each DNA restriction fragment (10 μ l or 1/3 of volume eluted from column) was ligated to 100 ng of pBluescriptII SK+ that had previously been digested with *Xba*I or *Xho*I. Ligation was mediated by T4 DNA ligase (Promega) at 16 °C overnight. The ligated plasmids (1/2 of ligation reaction) were transformed into chemically competent *E. coli* DH5 α . Positive CFUs were identified by α -complementation and grown for miniprep plasmid isolation. To confirm the ligation of the correct restriction fragment, the plasmids were digested with *Xba*I

or *Xho*I to release their insert. Separation of the digested plasmid DNA by agarose slab electrophoresis allowed sizing the insert fragment with calibrated DNA markers. Each plasmid was sequenced with the transposon-mediated Template Generation System (MJ Research).

PCR amplification of cDNA encoded by λ Zf13

Total RNA isolated from zebrafish whole bodies was used for cDNA synthesis as described in Chapter 2 for liver RNA isolation of liver RNA. The nested upstream primer set used for amplification were BreIIA.F (5'-CGC TAA AGA GAA ACG CCT CT-3') and BreIIB.F (5'-AAA GAG CCG TCG ATG GTT C-3'). These were used in combination with the downstream primer MCSRACE as described for 3' RACE. The PCR reaction mix was: 10X EnzOne 2000 buffer, 1.5 mM MgCl₂, 0.4 mM each primer, 1 U of EnzOne 2000 polymerase (ID Labs; London, Ontario, Canada) and 1 μ l of cDNA in a 50 μ l final reaction volume. The cycling parameters were: (1) 94 °C, 3 min; (2) 94 °C, 15 sec; (3) 65 °C, 15 sec minus 0.5 °C per cycle (4) 72 °C, 3 min (repeat steps 2 through 4 for 19 cycles); (5) 94 °C, 15 sec; (6) 55 °C, 15 sec; (7) 72 °C, 3 min (repeat steps 5 through 7 for 19 cycles); (8) fill-in at 72 °C for 5 min. The first reaction used the primer combination BreIIA.F/MCSRACE, and the second nested reaction used BreIIB.F/MCSRACE with 1 μ l of the first reaction as template.

Genome sequence database searches

Scaffolds resulting from whole genome shotgun assemblies were retrieved from Zv2 (www.ensembl.org) for *D. rerio*, JGI v3.0 (genome.jgi-psf.org/fugu6/fugu6.home.html) for *Fugu rubripes*, and Genoscope assembly v6 (www.genoscope.cns.fr) for *Tetraodon nigroviridis*.

Global alignments and comparative search for regulatory regions

Global DNA alignments and the search for conserved transcription factor binding motifs were performed using rVISTA [329].

Results and Discussion

A domain divided

In contrast to other fish genomes, the relatively small genome size of striped bass (825 Mb) [330] allows the use of conventional small insert bacteriophage libraries to screen for genes. Following high stringency washes and three rounds of screening, 16 positive plaques were isolated and analyzed by restriction digest. Small-scale preparations of bacteriophage DNA were digested with two enzymes, *EcoRI* and *SacI*, both with unique cut sites present in the polylinker regions of the bacteriophage vector. From the digest analysis, seven unique restriction patterns were identified that did not share bands, and possibly represented unique genes (fig. 59). A representative of each banding pattern was chosen for further analysis (λ Ms2, 3, 4, 7, 8, 10 and 15). The ends of the clone inserts were also sequenced to confirm their unique ends.

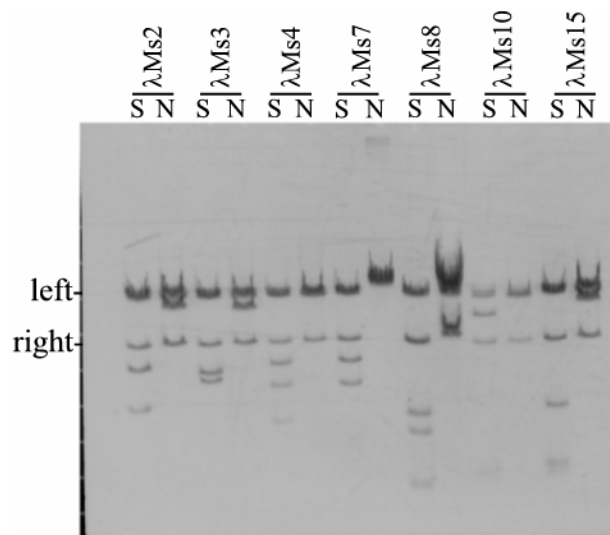


Figure 59. Restriction enzyme mapping of striped bass' genomic DNA clones. Bands corresponding to λ FIXII bacteriophage vector arms are labeled left (20.2 kb) and right (9.14 kb). Gel: 0.7% agarose, 1X TAE. Restriction enzymes: Sc=*Sac*I and N=*Not*I.

Subsequent attention was focused on one clone, λ Ms15, due to the stronger autoradiography signal it generated during Southern analysis of the restriction digest (fig. 60). Initially, subcloning was performed by PCR using λ Ms15 as DNA template and a gene specific primer and λ arm primer. The position of the putative MsaFBP32 gene was mapped with respect to the bacteriophage vector arms by PCR using exon specific primers and phage arm primers (bacteriophage RNA polymerase promoters T7 and T3). After several trials with the primers initially used for amplifying the MsaFBP32 cDNA, the largest amplicon of approximately 4.5 kb was obtained using the 5'UTR consensus primer, FBP1.F and the vector arm primer, HiT3. The amplicon was subcloned by TA cloning (pMs15s.1), and the insert was sequenced by ExoIII/S1 nuclease directional deletion sequencing (fig. 61).

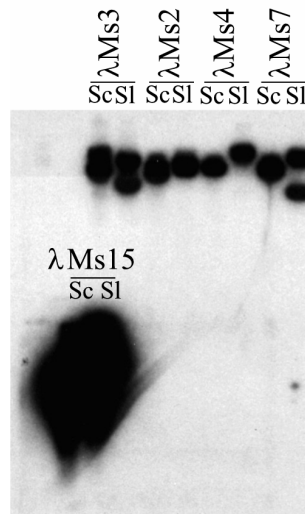


Figure 60. Autoradiogram of restriction analysis of λ Ms bacteriophage clones. Restriction enzymes: Sc, *SacI* and SI, *SalI*.

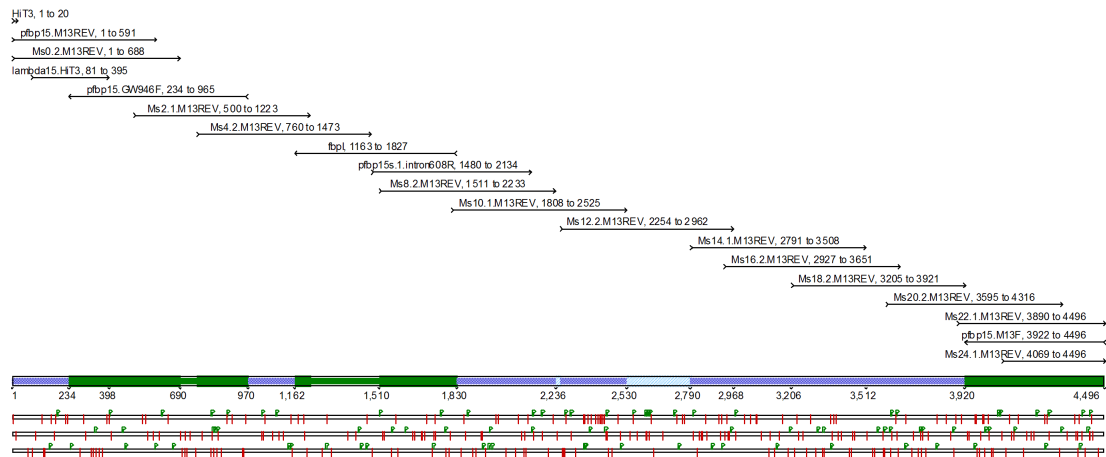


Figure 61. Sequencing contig path of directionally deleted subclones of pMs15s.1

The result was a sequencing contig path of 4.5 kb, which contains the full gene sequence of MsaFBP32 in addition to 847 bp downstream of the mRNA polyadenylation signal. A comparison of the cDNA sequence with the genomic clone sequence revealed the exon organization. The compact MsaFBP32s gene (3.5 kb) consists of six exons of which the last five encode the ORF. In detail, the first exon delimits the 5'UTR and the second exon delimits the signal sequence plus the first three amino acids of the mature protein sequence. In

contrast to the partitions evident in these exons, the 3'UTR is continuous with the last coding exon. Interestingly, the tandem motifs are each encoded by two exons (N-FBPL: exon 3-4 and C-FBPL: exon 5-6) with intron splice sites in identical positions within each motif. The coding exon phases are exon 2 (2-1), exon 3 (1-2), exon 4 (2-1), exon 5 (1-2), exon 6 (2-poly(A)+ signal). A sequence similarity search with the full gene revealed the presence of a Tc1/*mariner*-like element (TLE) pseudogene in intron II. It lays contiguous to exon 3 but its transcriptional direction is oriented opposite to that of the lectin. Similar placement of such element within an intron has been described [331], although in this case the TLE is apparently non-functional. The ORF is truncated and the inverted target repeats characteristic of the cut-out, paste-in insertion mechanism of this transposon family (fig. 63) [332], are absent, which indicates this is a pseudogene from this transposon family.

To enable comparison of gene structures between species, the white bass FBP32 gene was also amplified from genomic DNA using upstream and downstream the primers that anneal to the first exon and last exon, respectively. After sequencing (fig. 64) it became apparent that intron II was substantially longer (2,037 bp) in this species. In contrast to the striped bass gene (3 kb) (fig. 62, Panel A), MchFBP32 is larger (~5 kb) (fig. 62, Panel B). The resulting amplicon was processed for sequencing identically as pMs15s.1 and the resulting clone named pMc.3. As for the rest of the gene, the exon structure was identical, and all other introns presented similar sizes to the MsaFBP32 gene. Fish intron size is much reduced in comparison to mammalian genes [333]; the compactness of the FBP32 gene follows this same trend. Interestingly, the Tc1/*mariner*-like sequence is also present in the same orientation and position as in the striped bass gene, which suggests that it was present prior to divergence of the two species. Qualitative comparison of rare genetic events, like transposon insertions, are increasingly

useful in resolving patterns of divergence where quantitative comparison of genes has failed [311, 334, 335]. It is possible that tracing the presence of the TLE in binary FBPLs in other teleosts will help in identifying other genes orthologous to those in moronids. The NH₂-terminal sequence of a fuclectin isolated from the serum of the European sea bass [336] is virtually identical to that of MsaFBP32 suggesting they are orthologous. Detection of the TLE in the gene of this FBPL may provide some indication of the evolution of this unique feature.

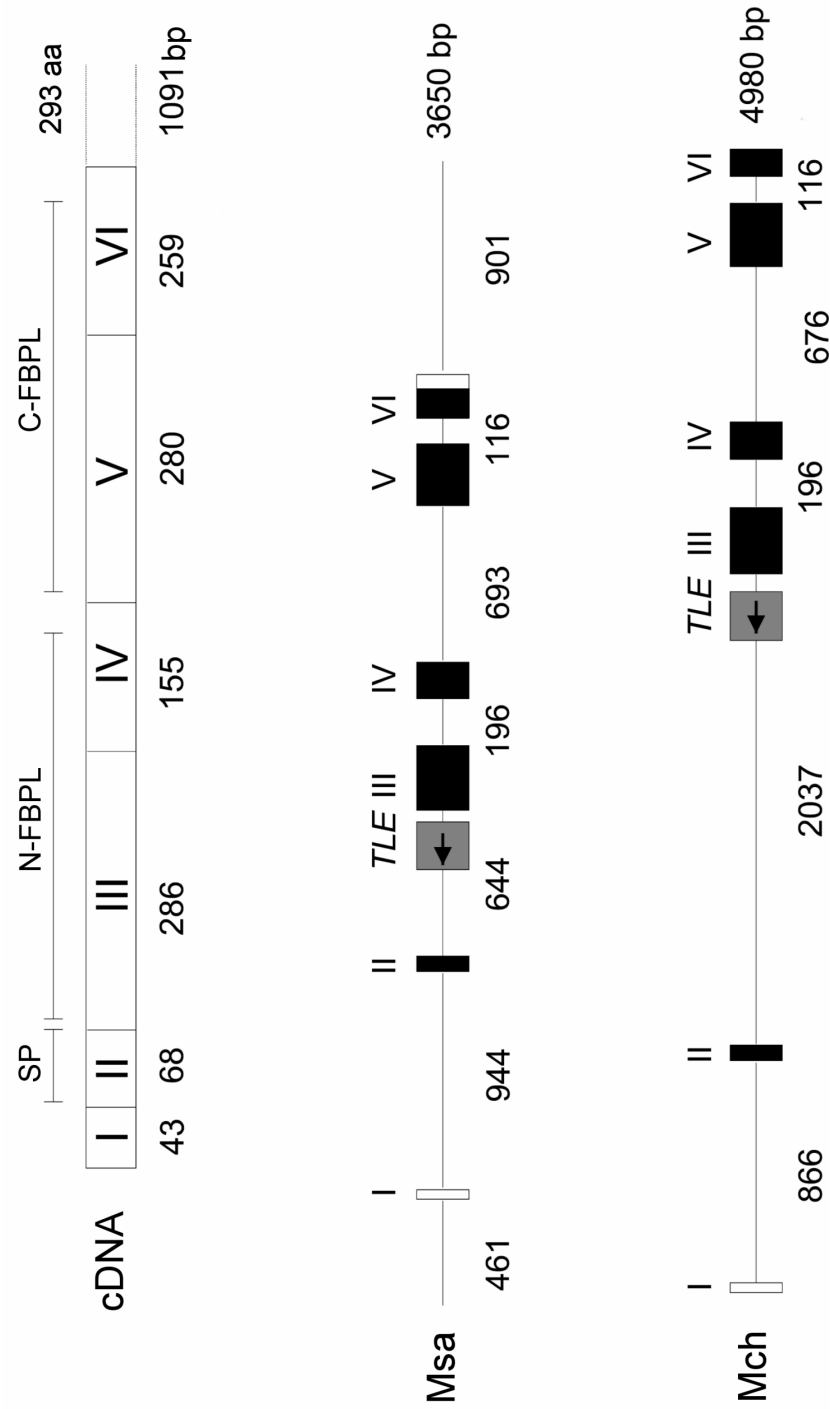


Figure 62. Structure of FBP32 cDNA and gene. **A.** Segments of FBP32 cDNA are labeled above according to encoded polypeptide motif: signal peptide (SP) and tandemly duplicated NH₂-terminal (N-FBPL) and COOH-terminal (C-FBPL) FBPL-like domains. Exon nucleotide lengths are placed below. **B.** Physical map of Msa and Mch FBP genes. Exons are enumerated with Roman numerals (above) and intron size is indicated below in bp. White boxes indicate non-translated exons while black boxes indicate translated exons. The Tc1/*mariner*-like element (TLE) pseudogene is indicated by the gray box with arrow pointing in direction of transcription.

	20	40	60	80	100
Mea-tce	---	---	---	---	---
Cel-tcl	VRPRLQAGHGEKVPKSEFKSKNEMVAKKHLRWGQFPAKXHL	---	---	---	---
Est	IRPR---	USPQSEDN-MAKRKK-FVAKKRLQNTNWKXKL	---	---	---
Csa-tss2	LSAHLKSGTLGVKVAQRPLSKKADSLGLFAKRLQNTNWKXKL	---	---	---	---
Dre-tdr2	ICSAHQSGTLGVKVAQRPLSKKADSLGLFAKRLQNTNWKXKL	---	---	---	---
Csa-tss1	VRKVIYRLNKGKSAKPLLOLPHKKKARLFAAGDKDRFTWRNVL	---	---	---	---
Elu	VRKVIYRLNKGKSAKPLLOLPHKKKARLFAAGDKDRFTWRNVL	---	---	---	---
Tal	VRKVIYRLNKGKSAKPLLOLPHKKKARLFAAGDKDRFTWRNVL	---	---	---	---
Cau	VRKVIYRLNKGKSAKPLLOLPHKKKARLFAAGDKDRFTWRNVL	---	---	---	---
Oca	VRKVIYRLNKGKSAKPLLOLPHKKKARLFAAGDKDRFTWRNVL	---	---	---	---
Ans	VRKVIYRLNKGKSAKPLLOLPHKKKARLFAAGDKDRFTWRNVL	---	---	---	---
Aku	VRKVIYRLNKGKSAKPLLOLPHKKKARLFAAGDKDRFTWRNVL	---	---	---	---
Dre-tdr1	VRKVIYRLNKGKSAKPLLOLPHKKKARLFAAGDKDRFTWRNVL	---	---	---	---
Rbe	VRKVIYRLNKGKSAKPLLOLPHKKKARLFAAGDKDRFTWRNVL	---	---	---	---
Cca	VRKVIYRLNKGKSAKPLLOLPHKKKARLFAAGDKDRFTWRNVL	---	---	---	---
Dal	VRKVIYRLNKGKSAKPLLOLPHKKKARLFAAGDKDRFTWRNVL	---	---	---	---
Ipu	VRKVIYRLNKGKSAKPLLOLPHKKKARLFAAGDKDRFTWRNVL	---	---	---	---

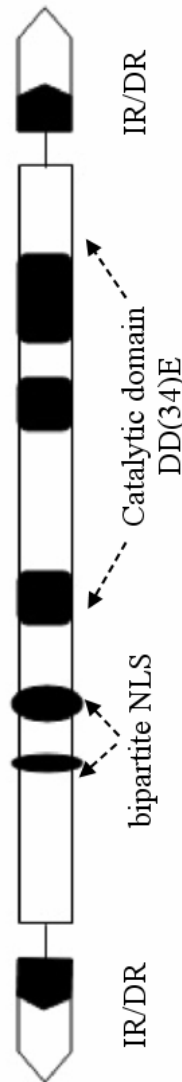
[illegible]

Figure 63. Tc1/*mariner*-like (TLE) transposons in FBP32 intron.

A. Multiple sequence alignment. The catalytic metal-coordinating residues, DD(34)E, are shaded in red. Species are aligned in the following order: Msa, striped bass; Cel, nematode; Est, Pacific hagfish; Ssa, North Atlantic salmon; Dre, zebrafish; Elu, Northern pike; Tal, white cloud mountainfish; Cau, goldfish; Ots, chinook salmon; Omy, steelhead trout; Aku, harlequin rasbora; Cca, common carp; Dal, pearl danio; Ipu, channel catfish. **B.** Scheme of prototype DDE transposon domain topology. IR/DR, inverted repeats/direct repeats; NLS, nuclear localization signal.

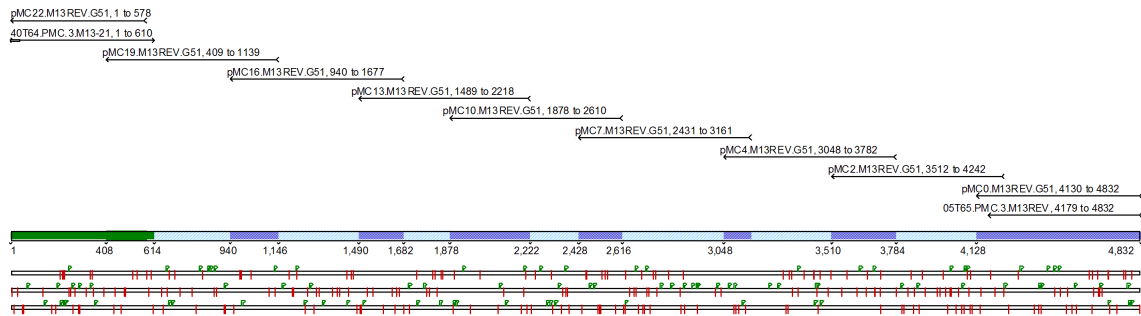


Figure 64. Contig path of directionally-deleted pMc3.

To test if both parental genes are present in the hybrid produced from crossing striped bass and white bass, PCR amplification was performed with upstream species-specific primers. Both genes were amplified from hybrid DNA while the parental species only produced amplicons with their respective primers (fig. 65). The outcomes for heterospecific hybridizations in fish are varied [337] sometimes excluding the male genetic contribution, but these results confirm that at least for this individual a hybrid genome was produced. Due to the high identity between the two species' gene products, it is difficult to determine by PCR if both genes are active in the hybrid, so the consequences that hybridization may have on expression of the parental lectins are presently unknown.

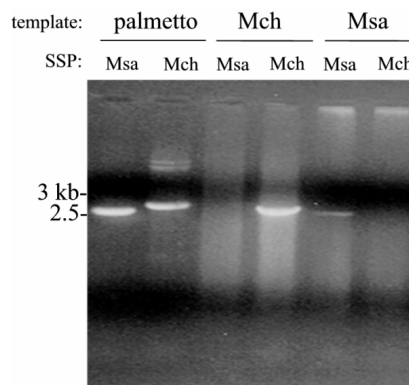


Figure 65. Detection in palmetto hybrid of both FBP32 genes from parental species. Mch, *M. chrysops*, Msa, *M. saxatilis*, SSP, species-specific primers.

The initial use of PCR for the purpose of gene walking yielded 380 bp upstream from the putative transcription start site. This region did not contain the prototypical TATA box or CAAT box common to many other genes [338]. The start site did show consensus though with the initiator (Inr) control motif (PyPyANT/A PyPy) that is present in some non-TATA containing promoter regions [338]. In order to investigate the presence of other possible regulatory motifs further upstream, additional sequence was sought. The λ Ms15 clone was known to contain in its ~15 kb insert a substantial length sequence upstream to MsaFBP32 gene (~11 kb). The full bacteriophage insert sequence was obtained from randomly selected subclones of the Sau3AI partially digested insert DNA. A combination of insert end sequences (i.e. shotgun sequencing) and transposon-mediated full subclone sequencing was used in building the sequence contig (fig. 66).

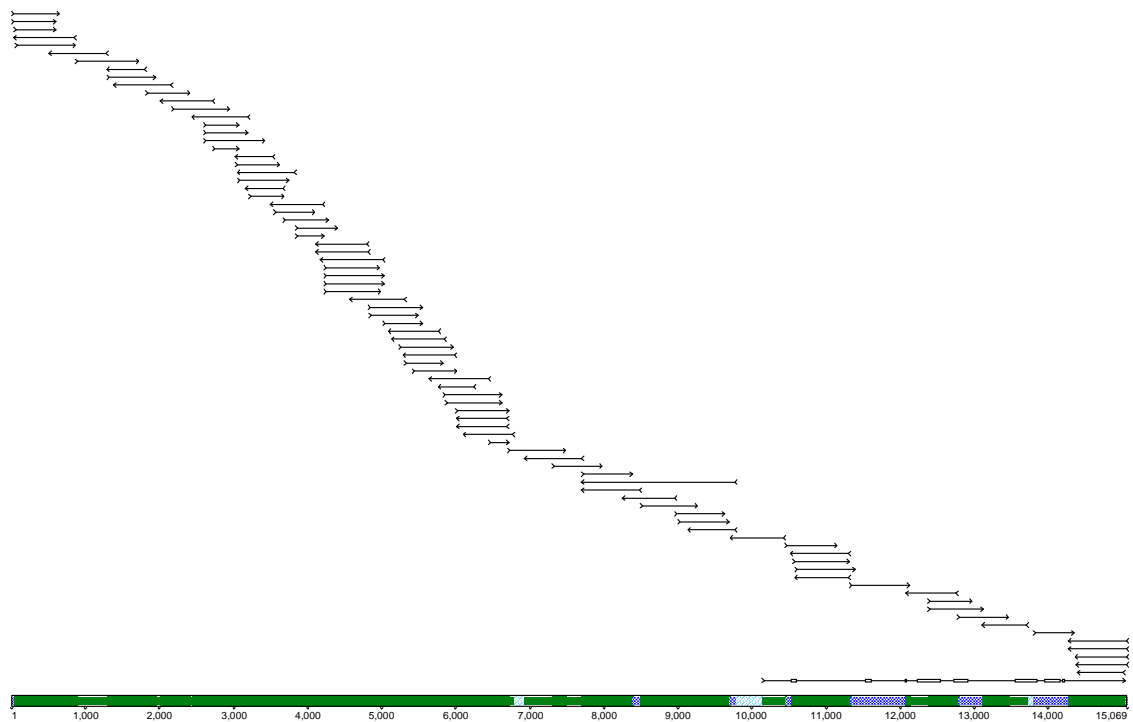


Figure 66. Sequencing contig of full λ Ms15 insert. The lowest arrow is that of pMs15s.1

Surprisingly, upon analysis of the completed sequence it was apparent that a duplicate MsaFBP32 gene was present upstream from the original (fig. 67). The putative product, hereon referred to as MsaFBP_{II}, also contains a duplicate FBPLs and a similar gene structure. The 3'UTR appears to be longer, as predicted from the most proximal polyadenylation signal. To confirm that, like MsaFBP32, the MsaFBP_{II} gene was expressed in liver, 3' RACE was performed with an exon-specific primer. A positive amplicon from liver RNA demonstrates that both genes are coexpressed but their relative abundance is presently unknown. The full cDNA sequence was completed by 5' RACE, allowing full annotation of the non-coding regions (fig. 68, Panel A). Comparison of the deduced polypeptide sequence indicates that they share 40% identity (70% similarity) indicative of a marked divergence (fig. 68, Panel B). The greatest differences not surprisingly lay in the UTRs, which commonly evolve more rapidly than the coding sequence.

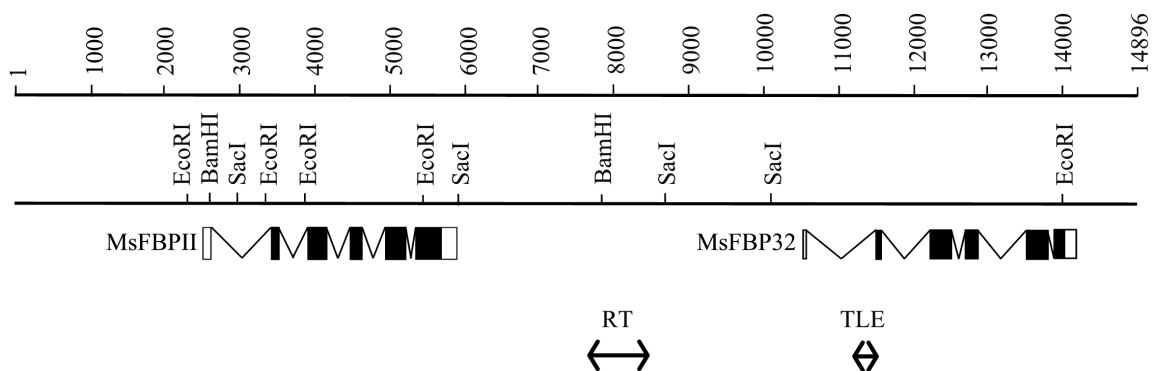


Figure 67. Map of λMs15 indicating the position of FBP genes and the transposon-like elements within. RT, inactive retrotransposon-like sequence; TLE, inactive Tc1/*mariner*-like element

The detection of a paralogous FBP sequence raised the possibility of family expansion within the bass genome. Although other genomic clones have not been sequenced, the fact that they hybridized with the FBP cDNA probe suggests that they contain an FBP-like sequence. One of these clones, λMs4, maps as contiguous to λMs15 (fig. 69), further supporting the

hypothesis that multiple tandem duplications have occurred for the FBP gene. The presence of repeats resulting from transposition events as annotated, could be conducive to the unequal crossing-over proposed to allow gene duplication.

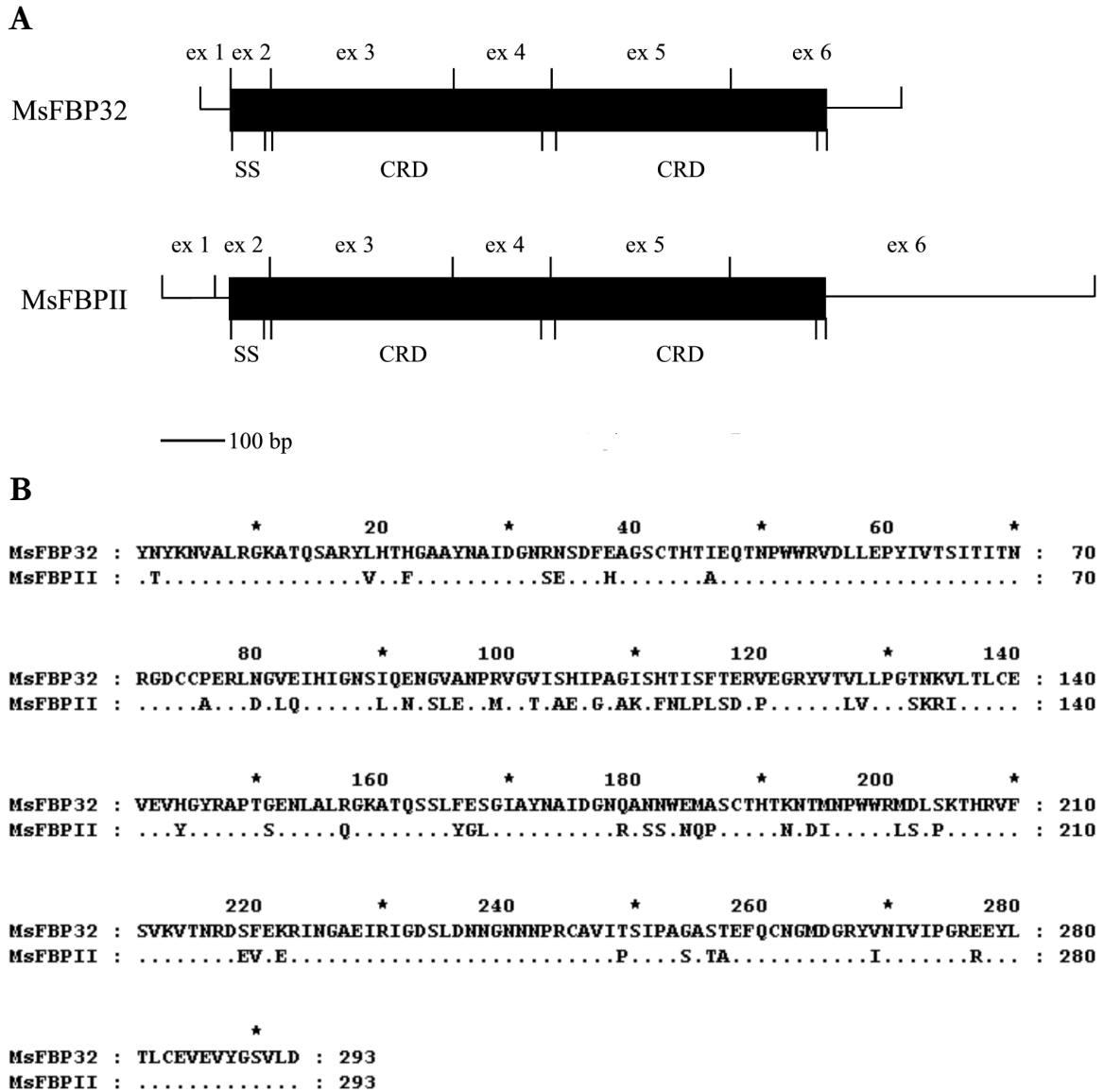


Figure 68. Comparison of MsaFBP cDNAs.

A. Schematic of exon positions and domains. Both drawings are at same scale. SS, signal sequence; CRD, FBPL carbohydrate recognition domain. **B.** Pairwise alignment of deduced polypeptide sequence. Identical residues are indicated with dots.

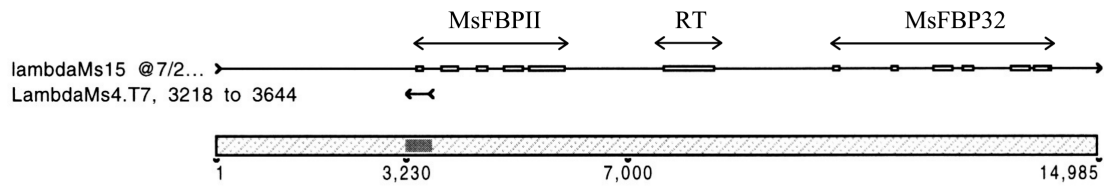


Figure 69. Overlap of λ Ms bacteriophage clones.

Exons of both MsaFBP32 and MsaFBP32 genes along with the retroposon pseudogene are indicated above the full sequence of λ Ms15. The overlapping end of λ Ms4 is shown by an arrow below λ Ms15.

As previously discussed, the ability for multiple FBP motifs to concatenate and to shuffle suggests they represent a structural domain unit. By resolving the gene structure, this issue can be better addressed. The presence of phase 1 introns delimiting each FBP motif supports this hypothesis, based on the trend observed for other domains commonly found as concatenated or mosaic proteins many of which are blood proteins [339]. Rarely are these shuffled domains flanked by introns of other phases. Conversely, the interruption of the FBP domain by a phase 2 intron may protect its integrity since splicing would require uncommon domains of that phase. Although the presented data illustrates gene organization in only a single fish genus, it may reflect the prototypical form for this lectin family. Further sequencing of FBP-like genes will help establish if the features observed for MsaFBP32 and MsaFBP32 are the norm, or if substantial differences remain to be described.

With the increased sequence information available from sequencing the full genomic clone a search was performed for transcription factor-binding site motifs (TFBS) [340]. The close position of the two FBP genes helped to focus the search to the intergenic region. The extensive matches to putative TFBSs were examined for other common core promoter sites (i.e. BRE, TATA Box, and DPE) but none was present. Core promoter structure can be varied, and in the absence of experimental promoter mapping studies, it is presently impossible to confidently annotate the regions regulating FBP expression. Eventually, with the

development of recombinant promoter/reporter constructs, the putative sites identified may be confirmed as relevant.

The completion of the fruit fly genome sequence [82] provided an opportunity to describe the gene structure for the sole FBP-containing receptors. Surprisingly, the second exon of the FBPL of CG9095 was fused with the exon encoding the CTL along with the second CCP domain (exon 4) (fig. 70). This is likely a derived state resulting from intron loss, since both CTLs [341] and CCPs [243] are typically flanked by phase 1 introns. This evidence, along with the fact that only two FBP-like domains are present in the fruit fly, suggests that this domain is no longer available for protein innovation. This likely occurred prior to the divergence of the fly from the mosquito, since both CG9095 and *furrowed* orthologues share this feature.

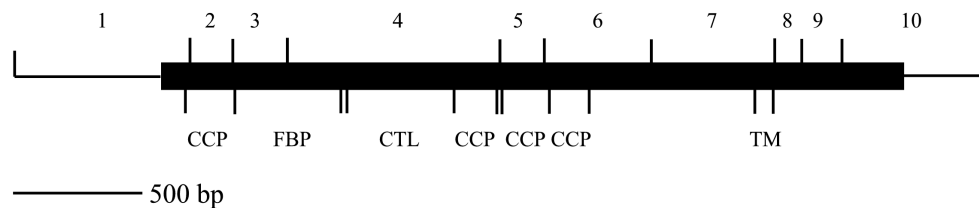


Figure 70. Scheme of the fruit fly CG9095 cDNA. The CDS is indicated by the filled rectangle. Intron splice sites and exon number are marked above. Protein domains are labeled below: complement control protein (CCP), FBP-type lectin-like (FTL), C-type lectin-like (CTL), transmembrane domain (TM).

Genomics of FBP in the zebrafish model

As discussed in chapter 2, amplification of zebrafish DNA with FBP-targeted degenerate primers resulted in the cloning of a cDNA that exhibited tandem domains and an extension at the COOH-terminus. Since this feature is unique to this protein, the gene structures determined for MsaFBP are of limited use. Therefore, a λ genomic library from zebrafish was screened with a partial DreI-FBPL cDNA probe available at that stage of the studies. Two promising clones were further characterized by the same procedure used for the striped bass,

but were only partially sequenced. One clone, λ Zf10, encoded from the C-FBPL downstream providing the exon structure of the COOH-terminal extension of interest (fig. 71). Since introns are typically short in fish genes, the intron lengths leading to the extension exons were substantially longer than expected.

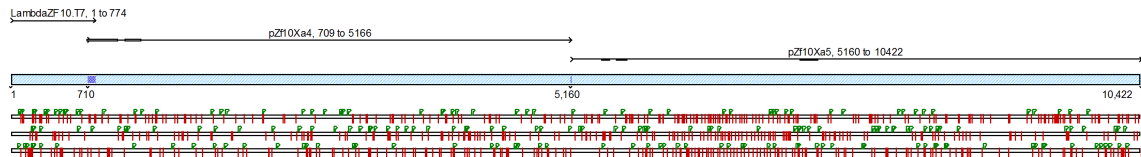


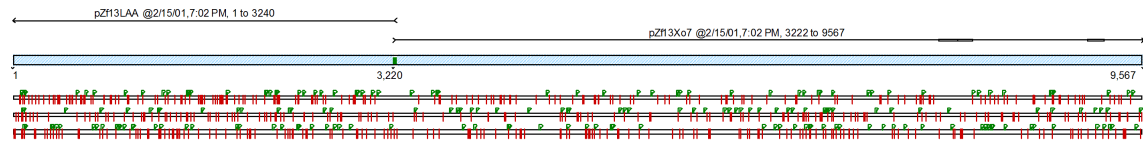
Figure 71. Cloning of DreI-FBPL gene.

Contig illustrates partial sequencing of λ Zf10 genomic clone insert. Boxes indicate exons confirmed by alignment to cDNA. The third and fourth exons encode for the C-terminal extension while the last exon encodes the 3'UTR. Leftmost arrow represents sequence obtained from the T7 end of cloning site.

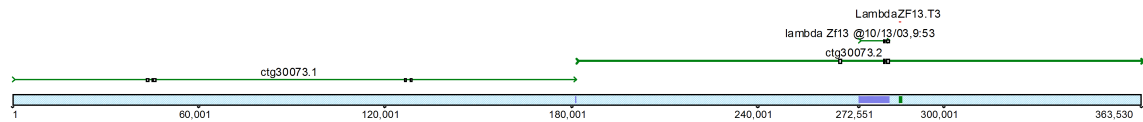
A second clone, λ Zf13, also encoded a single FBPL but despite substantial upstream sequence obtained (8 kb); no additional FBP domain was detected (fig. 72, Panel A). To help annotate the genomic sequence, RACE was performed with whole body cDNA in order to determine the complete transcript sequence. A successful 3' RACE indicated that following the FBPL there is a similar but unique extension, as observed in DreI-FBPL. Unfortunately, the 5' RACE did not succeed, and it remains to be determined whether this represents a single domain FBPL, as present in the Japanese eel. The rapid progress of assembling the whole genome shotgun sequence of zebrafish allowed searching for matches to these incomplete genes. No match is presently found for λ Zf10, but an extensive scaffold is available for λ Zf13 (fig. 72, Panel B). After comparing the protein sequence of the second homologue, DreII-FBPL, to the translation of this scaffold it is evident that at least two other FBPLs are confidently identifiable (fig. 72, Panel C). Closer upstream to the DreII-FBPL, the second exon encoding an FBPL appears to be present, which suggests that it may not be a single FBPL except for the absence of the first exon of this putative FBPL. Until a full transcript is

obtained, the exact sequence of this protein will remain unclear. It is evident from this scaffold, that as proposed for the striped bass, gene duplication led to an expansion of syntenic FBPLs.

A



B



C

Score = 105 bits (262), Expect = 4e-21Identities = 50/93 (53%), Positives = 65/93 (69%) Frame = +2

```

DreII: 1      NLALKRNASQSSSFGFWSAERAVDGSRTPGPKLWSVCSSTANQSNPWWRVLDGVDYRVSRV 60
              NLA +N QSS++ W E+A+D +          CSST Q++PWR+DLG +Y+VS V
Ctg30073.1: 128039 NLATGKNVMQSSSTYSSWIPEQAIDFNPLSDPSIGCSSTNGQTDPPWRLDLGHIVQVSTV 128218

DreII: 61     IITNINNSVADRINGAQIHIGNSLENNGTNNPM 93
              ++TN N +RINGA+IHIGNSLENNG NNPM
Ctg30073.1: 128219 VVTNRLNCCPERINGAEIHIGNSLENNGNNPM 128317

Score = 95.5 bits (236), Expect = 4e-18Identities = 48/86 (55%), Positives = 63/86 (72%) Frame = +1

DreII: 8      ASQSSSFGFWSAERAVDGSRTPGPKLWSVCSSTANQSNPWWRVLDGVDYRVSRVIITNINN 67
              A+QSS+ WSA+RA+DG R ++ CSST N+++PW RVDL Y V+RV+ITN N
Ctg30073.1: 45568 ATQSSSTSKDWSADRAIDGDRGLQQINKGCSSTLNETSPWLRVDLLYFYAVNRVITNRNV 45747

DreII: 68     SVADRINGAQIHIGNSLENNGTNNPM 93
              S A ++ G +IHIG+SLENNG NNPM
Ctg30073.1: 45748 SNAIQMTGLEIHIGSSLENNGNNPM 45825

Score = 48.9 bits (115), Expect = 4e-04Identities = 24/48 (50%), Positives = 33/48 (68%) Frame = +3

DreII: 94     CAVISSIPAGVSATFLCCFMGGRYVSLFIPGDSKMLTLCEVEVYVEGP 141
              CAV+S+IPAG + ++ C M+GRYV + I S +LTLC V V+V P
Ctg30073.1: 126207 CAVVSTIPAGQTFYSYSCNGMQGRYVFVDINAPSSILTLCVGVFVVP 126350

Score = 43.5 bits (101), Expect = 0.019Identities = 32/92 (34%), Positives = 44/92 (47%) Frame = +3

DreII: 1      NLALKRNASQSSSFGFWSAERAVDGSRTPGPKLWSVCSSTANQSNPWWRVLDGVDYRVSRV 60
              N AL R A +S G A A+DG L C+ T Q +PW R++L + YR++ V
Ctg30073.1: 43455 NALWRTAVLASQTGSSYARNALDG-----LPQTCAQTTFQDPDWIRLLNLEYRINVV 43616

DreII: 61     IITNINNSVADRINGAQIHIGNSLENNGTNNP 92
              + N N +NG IGN + NNP
Ctg30073.1: 43617 TMINSNLGYVPLNGTTTRIGND-PSYAYNNP 43709

Score = 42.4 bits (98), Expect = 0.042Identities = 27/81 (33%), Positives = 43/81 (52%), Gaps = 8/81 (9%) Frame = -1

DreII: 18     SAERAVDGSRTPGPK-----LWSVCSSTANQSNPWWRVLDGVDYRVSRVIITNINNSV 69
              SA+ + GSRT P L + C+T ++ PW +VDL VY+VS +T +
Ctg30073.1: 62127 SAQSVLCGSRTPMDGGKNFSTLSTCAATTWENFPWLKVDLQAVYQVSTATVITYREDRF 61948

DreII: 70     ADRINGAQIHIGNSLENNGTN 90
              D+ ++ G+SLE +G N
Ctg30073.1: 61947 PDKT--VTNFGSSLEFDGYN 61891

Score = 40.8 bits (94), Expect = 0.12Identities = 19/44 (43%), Positives = 25/44 (56%) Frame = +2

DreII: 94     CAVISSIPAGVSATFLCCFMGGRYVSLFIPGDSKMLTLCEVEVY 137
              C I + S F C M GRYV + +PG +LTLC+E+Y
Ctg30073.1: 45059 CTKIPNASPVTSTNFGSGMVGRYVFVHVPGYMAILTLCLEIY 45190

Score = 36.6 bits (83), Expect = 2.3Identities = 18/42 (42%), Positives = 28/42 (65%) Frame = -1

DreII: 158    LSEVQFLTQLESALARGLSDVTLHWTLQPKQMEQVAPAQR 199
              ++E L QL SAL RG+S++TL WT+ P++ EQ A ++
Ctg30073.1: 53613 MNEYLILFQLTSALTERGISNMTLSWTKTPEE-EEQTAAEEK 53491

Score = 194 bits (492), Expect = 9e-48Identities = 93/93 (100%), Positives = 93/93 (100%) Frame = +3

DreII: 1      NLALKRNASQSSSFGFWSAERAVDGSRTPGPKLWSVCSSTANQSNPWWRVLDGVDYRVSRV 60
              NLALKRNASQSSSFGFWSAERAVDGSRTPGPKLWSVCSSTANQSNPWWRVLDGVDYRVSRV
Ctg30073.2: 98070 NLALKRNASQSSSFGFWSAERAVDGSRTPGPKLWSVCSSTANQSNPWWRVLDGVDYRVSRV 98249

DreII: 61     IITNINNSVADRINGAQIHIGNSLENNGTNNPM 93
              IITNINNSVADRINGAQIHIGNSLENNGTNNPM
Ctg30073.2: 98250 IITNINNSVADRINGAQIHIGNSLENNGTNNPM 98348

Score = 99.4 bits (246), Expect = 3e-19Identities = 47/47 (100%), Positives = 47/47 (100%) Frame = +2

DreII: 94     CAVISSIPAGVSATFLCCFMGGRYVSLFIPGDSKMLTLCEVEVYVEG 140
              CAVISSIPAGVSATFLCCFMGGRYVSLFIPGDSKMLTLCEVEVYVEG
Ctg30073.2: 99371 CAVISSIPAGVSATFLCCFMGGRYVSLFIPGDSKMLTLCEVEVYVEG 99511

Score = 99.4 bits (246), Expect = 3e-19Identities = 47/47 (100%), Positives = 47/47 (100%) Frame = +1

DreII: 94     CAVISSIPAGVSATFLCCFMGGRYVSLFIPGDSKMLTLCEVEVYVEG 140
              CAVISSIPAGVSATFLCCFMGGRYVSLFIPGDSKMLTLCEVEVYVEG
Ctg30073.2: 85228 CAVISSIPAGVSATFLCCFMGGRYVSLFIPGDSKMLTLCEVEVYVEG 85368

```

Figure 72. Cloning of DreII-FBPL gene, a second homologue from zebrafish.

A. Subcloning of λ Zf13 genomic clone. **B.** Schematic of alignment of λ Zf13 partial sequence (shorter arrow at top) to zebrafish whole-genome shotgun contig ctg30073.2 (second assembly release Zv2) and upstream scaffold partner, ctg30073.1. Black boxes indicate segments encoding putative FBP-like exons. Numbering refers to nucleotide position. **C.** TblastN search of ctg30073 scaffold illustrating FBP-like matches distributed upstream of the DreII-FBPL locus. Note that the nucleotide numbering is independent for each ctg30073 contig.

A view of the whole fish

Among vertebrates, the second full genome to be sequenced after that of humans is the Japanese or tiger pufferfish (*F. rubripes*) [237], a member of Tetraodontids which is the most derived group of teleosts. Justification for sequencing this specie's genome emphasized its small genome size (400 Mb), and its potential contribution as a “genomic model” in the annotation of human genes [342-344] by facilitating identification of exon boundaries and gene regulatory regions [345]. Genome compactness is attributed to the reduction in intergenic DNA and intron size evidenced when comparing homologous chromosomal segments between the pufferfish and human [310, 346-350]. It is not clear why pufferfish have shed so much of their intergenic regions but one explanation initially offered was that it may correlate with the derived simplification of their body plan [330], although recent studies disagree [239]. A second genome-sequencing project for the freshwater spotted green pufferfish (*T. nigroviridis*), a smaller, more manageable species, was undertaken concurrently with the Japanese pufferfish [351-353]. It is estimated that both pufferfish species shared a common ancestor some 30 mya, as estimated from comparison of mitochondrial cytochrome b DNA sequences [351]. Although the genome has not been published, an initial sequence assembly is publicly available through the worldwide web.

With access to the first two full genome assemblies from teleosts, a unique opportunity arose for exhaustively searching for the full set of FBPL homologues in these species. After searching by TBLASTN the *F. rubripes* assembly (v2) stored at the U.S. Department of Energy's Joint Genome Institute, two scaffolds numbered 1786 (52.6 kb) and 5138 (15.7 kb) (fig. 74, Panel A) were identified as containing FBPL-like domains. A similar search was

performed for the *T. nigroviridis* genome (v6) which identified two more scaffolds, 794_4b (10.3 kb) and 461_3b (43.7 kb). All four scaffolds seemingly encode binary FBPLs similar to those of the striped bass and zebrafish. All other arrangements including single, higher order or chimeric were not found in either pufferfish genome. Neither genome had been annotated at that time. Presently, only the *F. rubripes* genome has been processed through an automated annotation pipeline. Therefore, each FBP gene was examined and gene structure was assigned manually following the features of MsaFBP structure. It must be stated since it is not illustrated that many gaps of unknown length are present in each scaffold as they represent the early stages of assembly. The identification of the coding exons was straightforward, since they exhibited two exons per FBPL, as previously described. Annotating the 5' UTR and the signal peptide of the gene was more difficult, since they apparently are always encoded by individual exons. This hinders the identification of the transcriptions start site (TSS) notoriously difficult to predict automatically. Fortunately, upon searching the dbEST with the deduced *F. rubripes* transcript from scaffold_1786, a matching 5' EST was identified (GenBank Accession AL842404.1) (fig. 73, Panel A). Interestingly, the provenance stated for this clone was from a heart cDNA library, a tissue previously unknown to express an FBP. Alignment of the EST to the scaffold (fig. 73, Panel B) allowed demarcating with certainty the 5'UTR exon, the signal peptide exon and first exon of the mature protein (Table 6). Unfortunately, the 3' EST of this clone is unavailable for annotating the 3' UTR, but the presence of a stop on the last coding exon prior to a splice donor consensus suggests that the CDS is complete. One caveat about this 5'EST sequence is that it appears chimeric as judged by the 5' end not being present in the extensive upstream sequence available in the scaffold; it does though match a

different scaffold. No EST was detected for the FBP present on scaffold_5138, so its 5'UTR remains unidentified and only a tentative signal peptide exon was demarcated (Table 6).

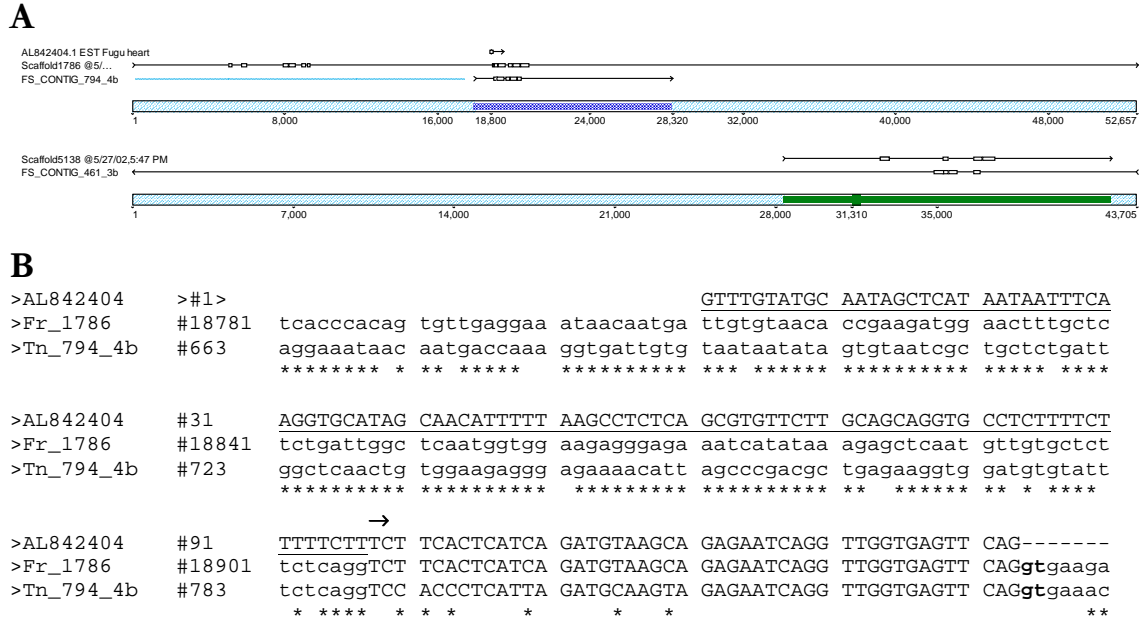


Figure 73. Identification of two pairs of orthologous FBP genes in pufferfish. **A.** Schematic alignment of orthologous scaffolds. Small arrow represents *F. rubripes* heart 5' EST. Boxes indicate FBP exons. See text for description scaffold provenance. Note the difference in lengths between the two species. **B.** Mapping of the TSS by alignment of *F. rubripes* heart 5' EST (AL842404) to the orthologous pufferfish scaffolds, Fr_1786 and Tn_794_4b. The first exon is indicated in uppercase nucleotides. The predicted splice donor site (**gt**) is indicated in bold. Dashes (-) represent gaps and asterisks (*) indicate dissimilarities. Note. The 5' half of the EST sequence (underlined) does not match any region within the scaffolds.

Table 6. Predicted intron splice sites of *F. rubripes* FBPL genes. Letters in bold are the canonical intron splice sites. N-FBPL and C-FBPL refer to NH₂-terminal and COOH-terminal domains, respectively. Complete cDNA sequences are not available for these genes, so non-coding exons are not reliably predicted and the size stated for the last exon is relative to the first polyA⁺ signal motif encountered. n.a.: not applicable; n/a: not-available

Exon	cDNA domain encoded	Exon size (bp)	Donor Site (exon intron)	Intron Phase	Acceptor Site (intron exon)	Intron size (bp)
scaffold_1786						
1	5' UTR	47		n.a.		86
1	Signal	71	CAG GTGAAGATTGCATTCAA	1	TGTTTTCTTTTAAACA AG TTT	117
2	N-FBPL	286	ATC GTGAGTAAAAAGGTTT	2	TTCTGTCTGCCTGAC AG AAA	242
3	N-FBPL	149	AAT GTGAGTGAAAAAATATA	1	TGTTCTCGCTCCTCT AG TGT	87
4	C-FBPL	277	CAG GTGACAAACCCCAACGT	2	CGTTGTTTCTGCA GGAG AGA	131
5	C-FBPL	375	CAG GTAGCCCACTCGATGCT	n/a	TTATTCTTTGTCC GTCA ATG	
6	+3' UTR(?)					
scaffold_5138						
1	5' UTR			n.a.		n/a
2	Signal	286		n/a	n/a	n/a
3	N-FBPL	146	CAG GTGAAGCGCCATTTCGAG	2	TTCTGTATGTCAATA AG GAC	74
4	N-FBPL	295	TTG GTATGGCCCTCACCGGC	1	TTTACTCCCTCAA ACAG GTG	1200
5	C-FBPL	559	CAG GTCTGTGGCAGGAGTGTA	2	GTTCTCTTTTCTTT TAG CAT	89
6	C-FBPL+3' UTR				CCGCTCCCCCGATT CCAG GTG	

Contrary to the case of the striped bass and zebrafish, the pufferfish possess only two copies of FBP genes, which tentatively appear not to be syntenic and distant from each other.

Phylogenetic analysis using their peptide alignment, as shown in chapter 3, suggests that the genes in *F. rubripes* 1786 and *T. nigroviridis* 794_4b scaffolds are an orthologous pair, as are *F. rubripes* 5138 to *T. nigroviridis* 461_3b.

Searching for the ON switch

Annotation of regions regulating gene expression [354] still remains a difficult endeavor due to the complexity of assembling transcriptional machinery [355], the degeneracy of transcription factor-binding sites [356] and their sometimes distant placement [357]. One approach that has shown promise is comparison of homologous genes, which share similar expression patterns. The selective pressure on the non-coding DNA relevant to transcription is expected to be preserved despite divergence. As discussed above, of the various core promoter motifs searched for around the TSS of the striped bass FBPs, only an Inr motif centered on the +1 position of MsaFBP32 was detected. Since both transcripts are coexpressed in liver, it is likely that their transcriptional control shares common features. Alignment of upstream regions including the 5'UTR (2 kb) was performed to search for consensus segments still conserved after gene duplication, but did not present substantial consensus between these genes (fig. 74). It therefore appears that control elements either are outside this region or are highly degenerate and not detectable by means of consensus. The paralogous relationship between these genes would suggest their regulatory regions might likely have diverged beyond recognition.

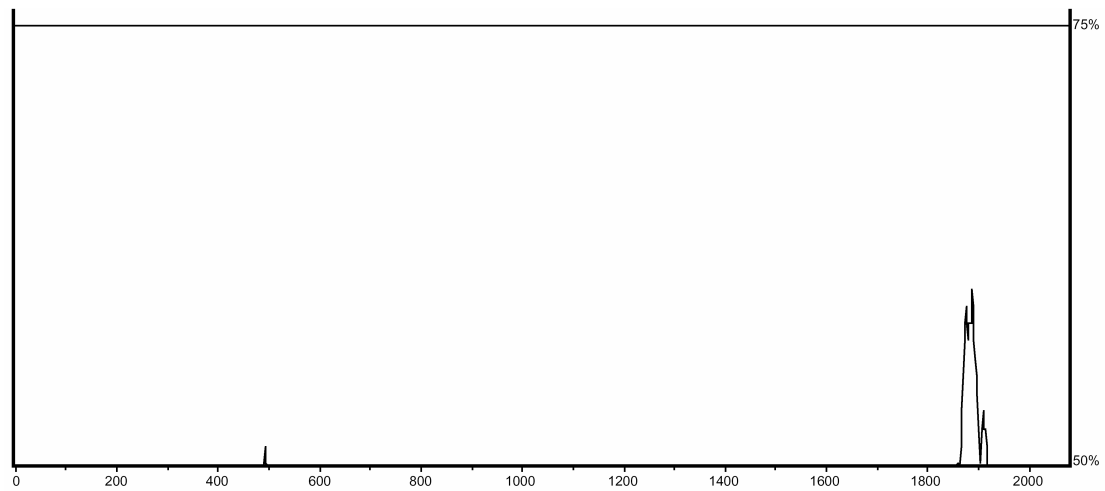


Figure 74. Alignment (rVISTA) of first exon and region upstream for MsaFBP11 and MsaFBP32. Histogram of percent identity (50-75%) vs. a 100 bp window moving along the alignment.

With greater number of genomes being completed, the study of genome evolution [328] seems more tractable, and has led to the development of publicly accessible tools for full genome comparisons [358]. Large-scale global alignments of widely diverging species have already produced promising results in finding potential gene regulatory regions [311]. In particular, teleosts which long ago diverged from tetrapods (450 mya) can be potentially useful for this approach [357, 359, 360]. Difficult phylogenies can also benefit from these alignments by identifying orthologues more confidently which share syntenic genes [361]. With the release of the *F. rubripes* assembly v.3, better quality scaffolds became available to implement this genome comparison approach. Previous scaffolds were elongated substantially relative to v.2: scaffold_1786 and scaffold_5138 were replaced by scaffold_599 (126.5 kb) and scaffold_237 (242 kb), respectively. Changes in accuracy are reflected in the reduction in the size of one unexpectedly long intron likely misassembled from the fragmentary sequence produced by the shotgun approach.

Using the rVISTA web site, the revised *F. rubripes* scaffolds were aligned to the unrevised *Tetraodon* scaffolds. In addition, putative transcription binding sites shared between species

were screened. The extensive consensus along the alignment of *F. rubripes* scaffold_237 and *T. nigroviridis* scaffold_461_3b strongly supports their orthology (fig. 75). The localization of the FBP gene well within the extensive length of each scaffold was favorable for searching for regulatory regions. Unfortunately, identification of the TSS for establishing where the genes begin has not been determined for these orthologues as mentioned above. Therefore, the search for core promoter regions could not be done with confidence. Still the alignment does illustrate consensus areas within both introns and flanking regions upstream and downstream that potentially regulate expression. Curiously, the intron dividing the FBPLs shows low consensus between the species due to a gap present in the *T. nigroviridis* scaffold. Since no further assembly releases have been made, it is presently unclear if it reflects a mutational event or is just a similar missassembly to that discussed for *F. rubripes*.

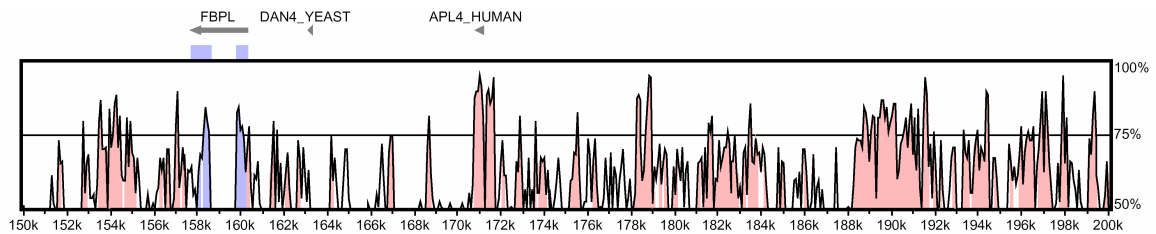


Figure 75. Global alignment of *F. rubripes* scaffold_237 to *T. nigroviridis* scaffold 461_3b. Histogram of percent identity vs. a 100 bp window moving along the alignment. Numbering refers to the longest scaffold (i.e. *F. rubripes* scaffold_237). Shading indicates $\geq 75\%$ identity within the window. Above the histogram, arrows indicate genes and boxes illustrate FBPLs.

Since the TSS of *F. rubripes* scaffold_599 was annotated with the help of an EST, it seemed a more promising source for describing the promoter regions. Comparison of *F. rubripes* scaffold_599 and *T. nigroviridis* scaffold 794_4b also presented a high percentage of consensi along the alignment proving their orthology (fig. 76, Panel A). Nevertheless, this alignment also failed to provide useful promoter sites due to the position of the FBP gene at the end of the short *T. nigroviridis* scaffold 794_4b. The quality of the sequence at the end of scaffold

794_4b is uncertain, although it is noteworthy that a rapid decline in sequence consensus occurs upstream of the TSS (fig. 76, Panel B).

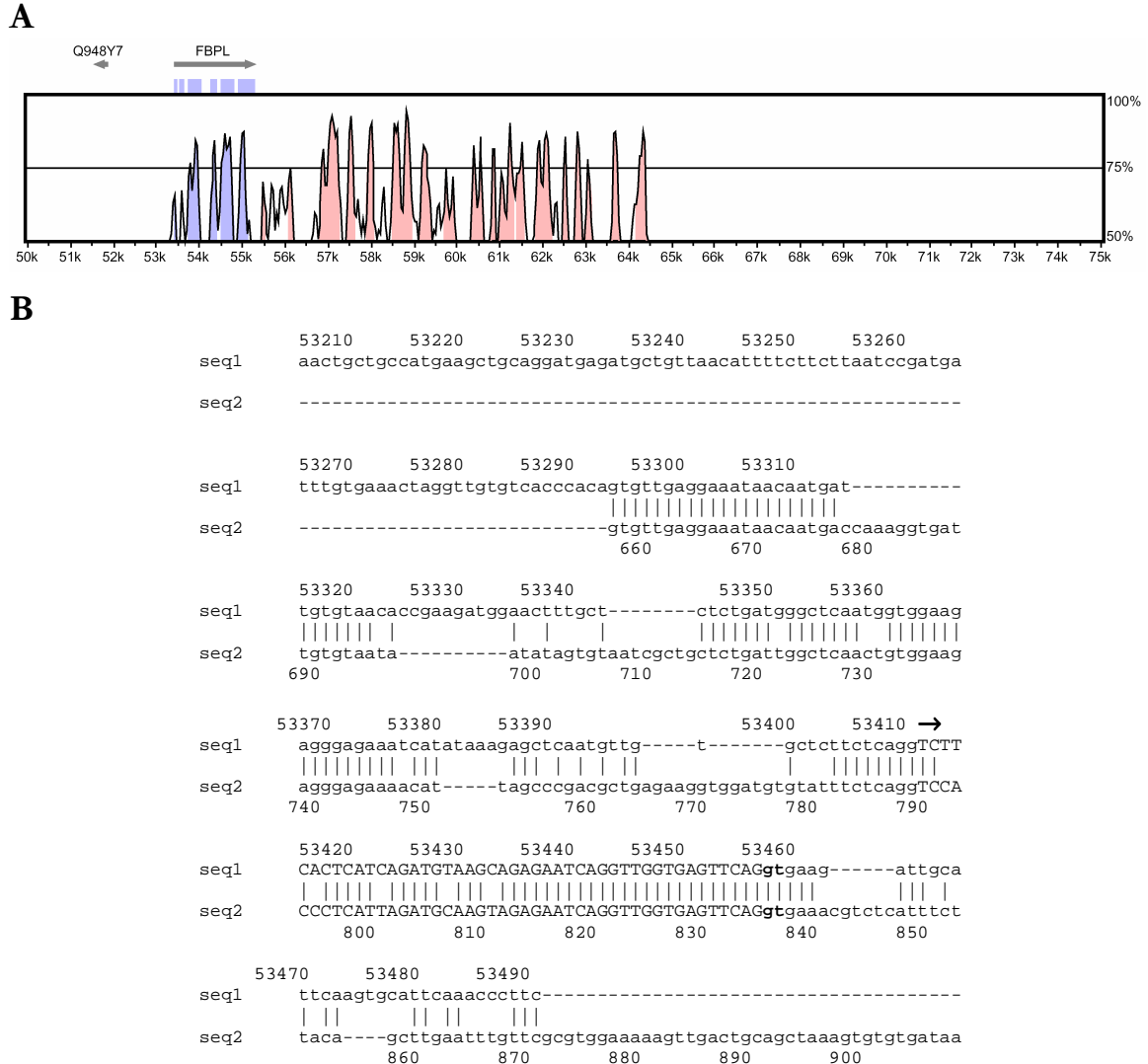


Figure 76. Global alignment of *F. rubripes* scaffold_599 to *T. nigroviridis* scaffold 794_4b. **A.** Histogram of percent identity vs. a 100 bp window moving along the alignment. Numbering refers to the longest scaffold (Fugu_599). Shading indicates $\geq 75\%$ identity within the window. Arrows indicate direction of gene transcription and boxes illustrate FBP exons. **B.** Detailed view of alignment at the TSS (arrow). The first exon is indicated in uppercase nucleotides. The predicted splice donor site (GT) is indicated in bold. Dashes (-) represent gaps and vertical bars (|) indicate identities. Seq1, *F. rubripes* scaffold_599; Seq2, *T. nigroviridis* scaffold_794_4b.

CHAPTER 5. STRUCTURAL ELUCIDATION OF AN FBPL

Heretofore, because the three-dimensional (3-D) structure of this lectin family has not been resolved, it was not possible to explain how FBPLs specifically recognize Fuc. Certainly, an alignment of the numerous FBPL family members facilitates gauging the permissibility of mutations within the primary structure. However, it is impossible to discern ligand-binding residues within the consensus residues, which more likely represent structurally important positions. For this reason, the process of mapping the FBPL binding site through a point-mutational approach could be a prolonged and tedious process and in the end inefficient since the spatial relationship of the residues in question would remain unresolved. Presently, X-ray crystallography and NMR spectroscopy remain the proven methods for macromolecular structure determination despite advances in *de novo* prediction [362]. Recent advances in protein *in vitro* expression, optimization in crystallization parameters, accelerated data collection with high energy synchrotron X-ray sources and computing capabilities have accelerated the speed with which new protein structures are presently being resolved [363]. Consequently, this has allowed much progress in understanding the structural diversity through which lectins recognize ligands [364-366]. Even a single resolved structure can greatly contribute to understanding the structure of other homologues [367]. This is because tertiary structure among homologues appears markedly conserved despite extensive mutational divergence from each other [368]. Based on this observation, large-scale efforts have been proposed to resolve representative structures for all protein sequence families not present in the structural database [369]. Thus, it was considered timely to attempt to resolve the first 3-D structure of an FBPL. In addition to MsaFBP32, the European eel (*Anguilla anguilla*) agglutinin (AAA) was included

in the crystallization trial because it was reasoned to be less flexible and therefore more amenable to forming a crystal since it lacks the spacer that separates the tandem domains present in MsaFBP32. For this, collaboration was initiated with a highly qualified, experienced crystallography team at the Department of Biochemistry and Biophysics at the Johns Hopkins School of Medicine, led by Dr. L. M. Amzel. In this chapter, the result of the elucidation of the structure for AAA bound with α -Fuc [370] will be discussed in an abridged form.

Materials and Methods

Repurification of AAA from a commercial preparation

Ten milligrams *A. anguilla* agglutinin (AAA) were purchased from Sigma (St. Louis, MO, USA). The lyophilized protein was first resuspended in 10 mM phosphate buffer (pH 7.2) and the soluble fraction was passed over a Fuc-Sepharose 6B matrix (1 ml) which was washed with 50 ml (50 column volumes) of phosphate buffer. The active lectin was eluted using 5 ml (5 column volumes) of 200 mM Fuc in phosphate buffer. Only half of the amount stated on the purchased vials (5 mg) was finally recovered after affinity chromatography. The substantial losses may be attributed to the insoluble particulate observed after resuspension.

Size exclusion chromatography

The native molecular weight of AAA was estimated by size exclusion chromatography as described in chapter 2.

Crystallization

Crystallization of AAA was performed by the hanging drop method. Briefly, the well from a 24-well polystyrene plate was filled with 1 ml of reservoir solution and covered with a silanized

round coverslip onto which a drop was spotted composed of equal volumes of AAA (10 mg/ml protein, 5 mM Fuc, 10 mM NaCl, 10 mM Tris-HCl (pH 7.5), and 1 mM CaCl₂) and 40-50% (v/v) 2-methyl-2,5-pentanediol (MPD) at pH 7-9 reservoir solution. Rhombohedral crystals of AAA grew to 0.3 mm in 72 h at 18 °C. Crystal dimensions (a=65.5 Å, b=65.5 Å and c=245.1 Å) indicates it belongs to the R32 space group. For the purpose of phase determination by isomorphous replacement, crystals were soaked overnight in reservoir solution containing 500 µM K₂AuCl₄.

X-ray diffraction data collection

Diffraction data was detected by a Rigaku Raxis IV⁺⁺ image plate (Danvers, MA, USA) from cryomounted (100 K) crystals bombarded with radiation from a Rigaku RU-H3R rotating CuK α anode. The resulting data was processed with Rigaku Crystal Clear software.

Structure determination

After detecting a single gold position from the Patterson difference map, refinement was initiated with MLPHARE from the CCP4 software suite [371]. Two additional gold positions were detected after the partial solution with the single site and were subsequently included in the refinement, which finally reached 1.9 Å. The resulting high-resolution electron density map facilitated building a structural model in spite of not knowing the primary sequence. The right-handedness consecutive β -chains indicated the enantiomer. Residue position identity was tentatively assigned based on the primary sequences of the Japanese eel fucosectin and positions where the density data disagreed a new residues was assigned based on the density shape and the protein environment. The coordinates and structure factors were deposited in the Protein Data Bank (PDB accession 1K12) (www.rcsb.org/pdb/).

Results and Discussion

Crystallization of AAA

The initial search for crystallizing conditions for AAA made use of 75 different solutions in a hanging drop format. Surprisingly, in a period of 48 hours usable crystals were obtained from a protein concentration of 5 mg/ml with MPD as a precipitant. The phase of the diffraction pattern was subsequently obtained from a gold-derived crystal. Within each rhombohedral unit cell, there were 18 asymmetric units representing AAA monomers. In an attempt to determine the character of the interaction of Fuc with AAA, the monosaccharide was included in the precipitating solutions in hopes it would co-crystallize with the protein. Successfully, the refined electron density map revealed one monosaccharide per peptide subunit, which illustrated the fucose to be solely in the α anomeric configuration. In total, the refined molecular model of AAA subunit includes 158 residues, 3 ligands (1 α -Fuc, 1 chlorine anion, and 1 cation (either Na^+ or Ca^{2+}), and 130 solvent molecules. Moreover, the final refined map revealed weak indications of alternative amino acid residues at some positions, which suggest the minor presence of other isolectins in the crystal. This observation suggested the presence of multiple isoforms in AAA, which may represent a situation analogous to that described for the agglutinin in the Japanese eel [182].

A novel lectin fold

The AAA structural model (fig. 77) confirms that the FBPL motif represents a folding domain as previously supposed. It is unique from the other lectin folds described so far [365, 366, 372-374] and therefore, was named F-type lectin domain (FTLD) following the convention previously established in the literature. Specifically, the fold is that of a β -barrel formed with a

jellyroll topology. The barrel consists of eight antiparallel β -strands arranged into two β -sheets almost sandwiched together. The first sheet consists of five β -strands ($\beta 2$, $\beta 3$, $\beta 10$, $\beta 6$ and $\beta 7$) and the second sheet of three β -strands ($\beta 11$, $\beta 5$ and $\beta 8$). The bottom of the barrel is covered by two short β -strands ($\beta 4$ and $\beta 9$). The NH_2 - and COOH - ends of the polypeptide form a two stranded β -sheet ($\beta 1$ and $\beta 11$), which enter and exit together from the closed end of the barrel structure, which may explain the domains tendency to concatenate. The sugar-binding site is located at the concave end of the barrel and is surrounded by five loops, which connect the principal β -sheets. Due to the sequence variability of homologous stretches evident in the sequence alignment these loops have been named CDR 1 through 5 in analogy to the complementarity-determining regions described for immunoglobulins. Surprisingly, one of these loops (CDR4: Arg79-Ala90) possesses the particular contiguous cysteines conserved among eukaryotic FBPLs. These reactive residues are exposed to the solvent and form a cystine, a feature that has not been described in other proteins, that closely interacts with the sugar ligand via a hydrophobic interaction. As for the other cysteine residues present in the FBPL motif, they also form the predicted disulfide bridges, which appear to stabilize the fold. The cystine formed by Cys50-Cys146 is apparently essential, as judged by its preservation in all eukaryotic family members, for stabilizing a loop rich in short 3_{10} helices and even participating (Cys146) in the coordination of the enveloped cation. In contrast, the disulfide bridge formed by Cys108-Cys124, which brings together strands $\beta 7$ and $\beta 8$ from opposite sheets, appears dispensable in some FBPLs. As mentioned above the ubiquity of cysteines in eukaryotic FBPLs markedly contrasts with their absence in prokaryotic FBPLs. It appears that *S. pneumoniae* and *M. degradans* FBPLs either do not require disulfide bridges for stability or reside

intracellularly where the high redox potential makes bridging impossible, therefore releasing selection for these residues. Intradomain interactions such as salt bridges also have a role in stabilizing the structure. One specific salt bridge between Arg41 and Glu149 attaches the loop between $\eta 2$ and $\eta 3$ helices to the $\beta 11$ strand. A second salt bridge between Asp64 and Arg131 creates an intrasheet link. The result of these features is a compact globular fold (43 Å l, 30 Å w) that is stabilized through disulfide bridges and metal coordination.

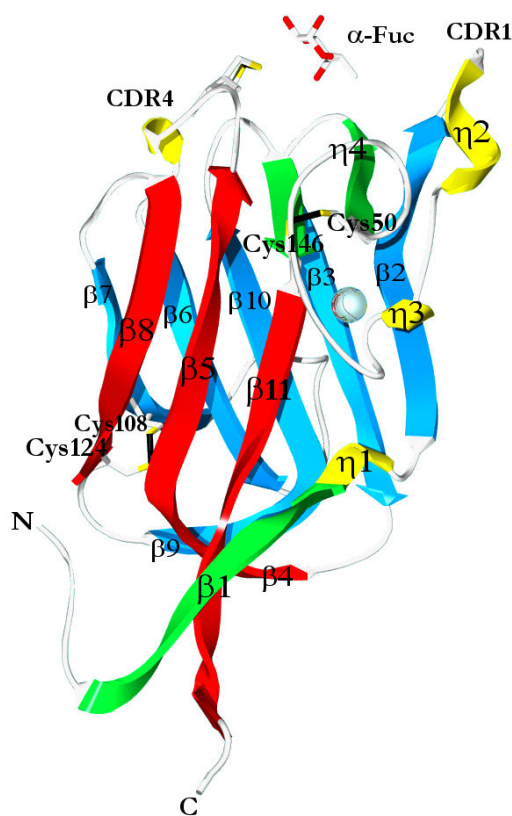


Figure 77. F-type lectin fold.

Ribbon diagram with helices colored yellow and β -sheets colored red (three-strand sheet) and blue (five-strand sheet). The β -strands in green form a 2-strand sheet. Bound α -Fuc is shown as a stick model above the lectin. Calcium is shown at right as a silver sphere.

Saccharide-binding site

The eel fucolectin had a notable role more than 50 years ago demonstrating that carbohydrates, specifically Fuc, and not proteins were the epitopes of the ABO blood group

[183]. Therefore, it was of great interest to identify the mechanism by which AAA bound Fuc. Although determining the 3D-shape of a protein itself contributes a wealth of detail, localizing a superficial binding-site can be difficult if the ligand is not present in the resolved structure [178, 253]. By co-crystallizing AAA with Fuc, the saccharide-binding site was unambiguously identified to the concave end of the barrel. As for other lectins a combination of molecular interactions explain the specificity demonstrated by AAA for Fuc. The monosaccharide is bound by a positively charged, shallow depression (fig. 78, Panel A) resulting from the close placement of a trio of basic side chains which contribute to binding by forming hydrogen bonds. However, no cations are directly involved in binding. From the base of the CDR 4 loop, two arginines (Arg79 and Arg86) donate hydrogens to the 3-, 4- hydroxyls, and the ring O5 positions (fig. 78, Panel B). The axial 4-hydroxyl is the only moiety that exhibits cooperative hydrogen bonding, a characteristic of protein-saccharide interactions, as both Arg79 and Arg86 donate to this oxygen while the N ϵ 2 of His52 accepts the hydroxyl hydrogen. Arg86 has a prominent role in orientating the saccharide ring as it donates to the vicinal 3- and 4-hydroxyls, which are at optimal distances by being equatorial and axial, respectively. The concentration of interactions at the 4-hydroxyl emphasizes the importance of the L- configuration of Fuc. In contrast, its enantiomer, D-Fuc (i.e. deoxy Gal), preferentially adopts a 4C_1 configuration, which would not present the preferred hydroxyl configuration. The orientation of the 2-hydroxyl away from the protein surface makes it a good candidate for further glycosidic polymerization. Specificity in lectins is frequently achieved through interactions with key features of the saccharide (i.e. hydroxyl disposition) that distinguish the favored ligand from the rest [50]. The disposition of the key feature is determined relative to common features (e.g. 3-hydroxyl and the ring oxygen) shared by all hexoses (i.e. Gal, Man and

Glc and their derivatives). Similarly, AAA orients Fuc using the 3-hydroxyl and the ring oxygen to distinguish the axial disposition of the 4-hydroxyl. This characteristic and the absence of a hydrogen acceptor in AAA for the 3-hydroxyl suggests that this moiety may also be suitable for glycoside extension, but steric hindrance may not allow this because of its proximity to the binding site. The placement of these basic residues is optimized through a network of hydrogen bonds from other residues and water molecules surrounding the binding site. Other interactions are also evident at the binding site such as van der Waals contacts between the ring C1 to C2 bond and the unusual cystine of CDR4. This close distance between CDR4 and Fuc would lead to steric clash if the anomeric configuration were β , which explains the selectivity for the α anomer since the C1-hydroxyl faces away from the protein surface. Contribution of the hydrophobic effect is evident in the nestling of deoxy C6 into a hydrophobic pocket where the Phe45 side chain lays, which explains the preference for Fuc over L-Gal. However, like most lectins, specificity is not absolute and new interactions may sometimes even arise with unnatural ligands leading to higher affinity [375]. In the case of AAA, an inverted 3-O-methyl-D-Gal can partly mimic the configuration of groups in Fuc, which may explain why it is a good inhibitor [376]. In summary, binding of AAA is mediated by few residues exploiting the δ -guanido moiety of Arg to form a lattice of hydrogen bonds with neighboring oxygens.

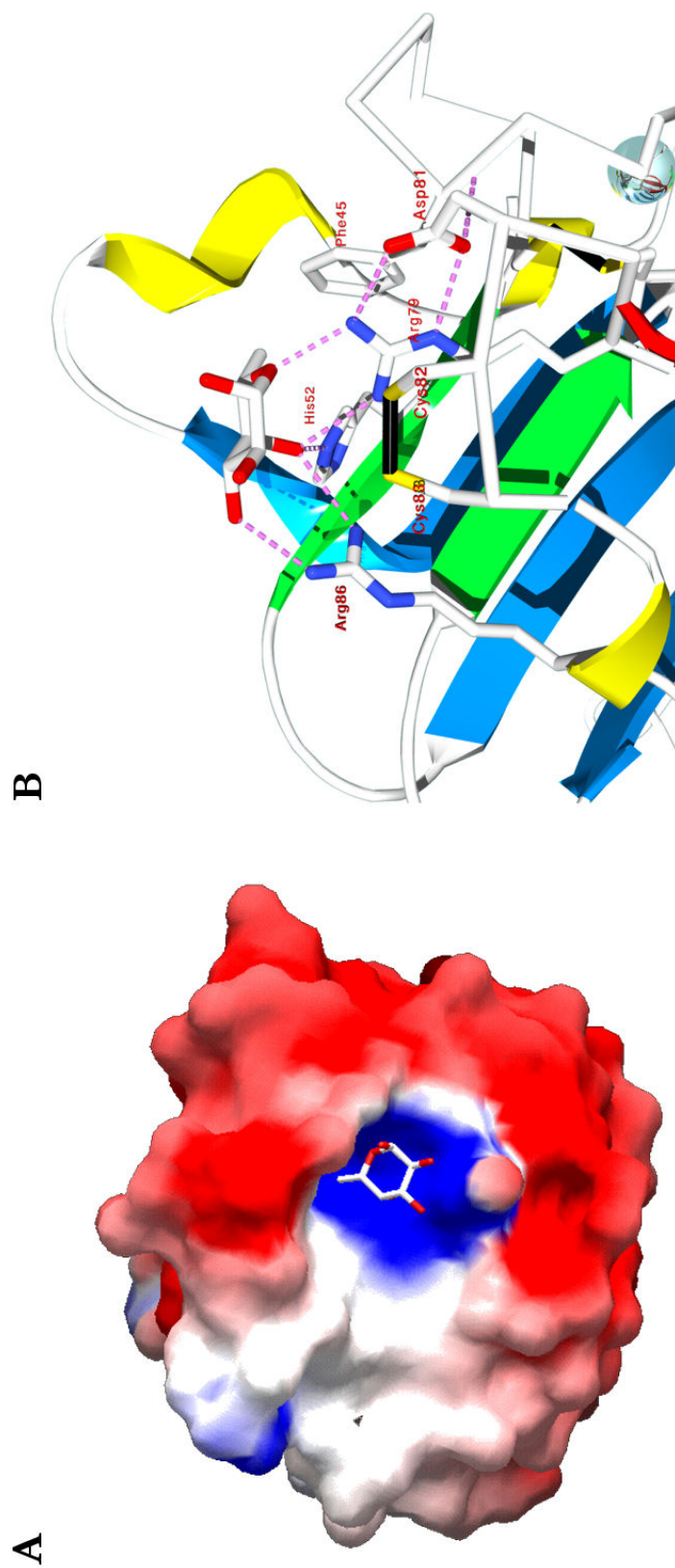


Figure 78. Sugar-binding site.

A. Electrostatic potential energy surface of AAA. The ligand-binding pocket with the bound sugar visible as a stick model is rich in basic residues evident by the blue coloration (> 10 kJ). Negative energy values (< -10 kJ) are colored red. **B.** View of α -Fuc bound to AAA. Broken pink lines represent the hydrogen bonds involved in ligand binding. Oxygen is colored red. The exocyclic C6 is proximal to the face of Phe 45, which provides a hydrophobic environment to accommodate this group. The silver sphere illustrates position of the bound metal relative to the carbohydrate-binding site.

As with MsaFBP32, AAA is present in circulation where it is more likely to encounter complex polysaccharides, such as oligosaccharides (i.e. blood group antigens) [377] or enterobacterial exopolysaccharides (i.e. colanic acid) [378] rather than free Fuc. These larger ligands are likely to form additional interactions to those observed with the monosaccharide alone. The identity of the natural ligand of AAA is presently unknown, but inhibition studies [379] with α -fucosylated oligosaccharides supports that additional interactions occur with subterminal saccharides. Reactivity of AAA to lactosamine-linked (α 1-4) Fuc (type 1: H-antigen and Lewis^a), but not to neo-lactosamine-linked (α 1-3) Fuc (type 2: Lewis^x) indicate the dramatic effect that just changes in linkage has on binding. Therefore, in addition to the primary site, surrounding sub-sites likely influence the binding of larger oligo- or polysaccharides. The interactions of AAA with fucosylated oligosaccharides, H-antigen type 1 and Lewis^a, were confidently modeled using the solution conformation of these oligosaccharides derived from conformational studies. The principal interactions are with polar residues from the rigid CDR1 loop, which rises above the binding shallow and with Tyr46 of CDR2. Interestingly, the CDR1 loop is variable among the several cDNAs cloned from the Japanese eel suggesting that it may regulate specificity for certain glycoconjugates. It is evident from the co-crystallization data and the models that the selectivity of the principal binding site of AAA is for α -Fuc; moreover, the variable surrounding loops have the potential to regulate binding of native glycoconjugates. The close interaction with the 3- and 4-hydroxyls, and C6 would preclude binding of substituted fucosyl-containing polysaccharides such as fucoidan [175] and colanic acid, which discounts a potentially relevant and widespread molecular pattern unique to bacteria.

Calcium-binding site

A conspicuous feature revealed by the AAA 3-D structure is a metal nestled within the loop rich in short 3_{10} helices (fig. 79). The geometry of the heptacoordinated site is pentagonal bipyramidal mediated by oxygens contributed by both the peptide backbone and side chains. In detail, the first coordination shell ligands are mediated by the semicontinuous residues Asn35 (O), Asp38 (O δ 1), Asn40 (O), Ser49 (O, O γ 1), Cys146 (O), and Glu147 (O ϵ 1). Despite only six residues being involved in coordination, heptacoordination is achieved by Ser49 performing as a bidentate ligand. One shortcoming of crystallographic data with regard to studying metal-binding sites is that it permits establishing only indirectly a metal ligand's identity. Fortunately, features such as coordination geometry and ligand-metal distance, which correlate with specific metals, reduces the possible candidates. Coordination distance is especially useful for identifying metals since it is rarely restrained during the structure refinement process producing accurate final measurements. Both the coordination distance (2.4-2.6 Å) and geometry present in AAA are consistent with parameters observed during crystallographic analysis of Na⁺ or Ca²⁺ coordination by small molecules. Selectivity for Ca²⁺ over other metals, including Na⁺, is achieved in proteins by coordination with neutral ligands [380] such as main-chain carbonyls and alcohol moieties. A survey of structures of known calcium-binding proteins [381] indicates that oxygen, frequently from the peptide carbonyls, is always the ligand at these sites. Comparison of coordination numbers points to seven as the norm but can vary from five to eight. Since AAA shares all these features, Ca²⁺ is most likely the identity of the cation. However, only chemical analysis of metal content of AAA can unequivocally identify the metal.

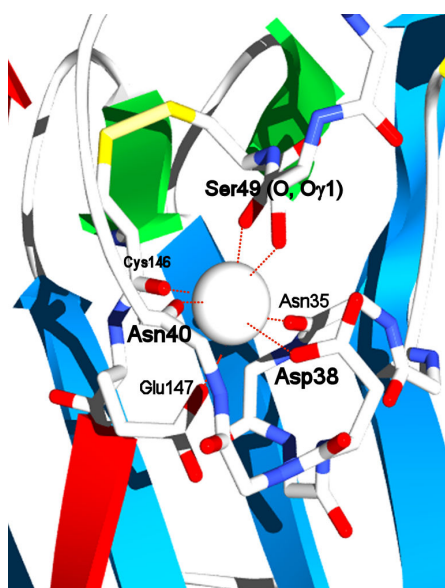


Figure 79. Heptacoordinated Ca^{2+} -binding site of AAA.

The principal role that calcium appears to play in AAA is structural stabilization. In contrast to C-type lectins, none of the metal-coordinated residues interacts with the carbohydrate ligand, therefore supporting solely a structure role. Intracellular levels of calcium are tightly regulated (100 nM) due to the importance of this cation as a secondary messenger in cellular signalling, but in the extracellular environment, a higher and relatively constant level (2.5 mM) is detected. Therefore, calcium-binding by proteins in the extracellular space is more related to their stability rather than to signaling [381]. The high coordination number of Ca^{2+} makes it an ideal crosslinker for bringing together different parts of a polypeptide acting in essence similar to a disulfide bridge. Previous studies of cation requirements in tachylectin-4 [181] and the Japanese eel fucoselectin [182] suggest that calcium is involved in sugar binding. In contrast, agglutination by MsaFBP32 was not significantly affected by chelating despite possessing similar residues at homologous positions. Although the evidence from the AAA structure rules out any direct relationship between cation and saccharide, it is possible that the removal of calcium induces changes in the CDR1 and CDR2 loops, which lead to abrogation of

binding. An explanation for the discrepancy between the cation requirements of two fish FBPLs is that calcium removal may actually affect quaternary structure rather than tertiary structure. Both tachylectin-4 and eel fucoselectin form larger order assemblies, which may be susceptible to chelation resulting in their disassembly. In contrast, the native monomers MsaFBP32 may remain active even after chelating since they possess two intramolecular FBPL domains, which could still allow agglutination. Further studies should address this possibility.

Quaternary structure of AAA

Unlike NMR spectroscopy, protein X-ray crystallography relies on the ability of dissolved polypeptides to assume an abnormal crystalline state. Inherent during the crystallization process is the constraint of molecular movement in the densely packed lattice. However, crystals from macromolecules are typically less dense and much intermolecular space remains filled by solvent. Consequently, the level of hydration of a crystallized protein is comparable to when it is dissolved. Therefore, in the absence of any deformations associated with crystal contacts, the crystallographic structure confidently reflects the native shape of the protein under physiological conditions. In the crystallographic model of AAA, the quaternary structure appeared as a homo-trimer of cyclic symmetry (i.e. C_3 point group) (fig. 80, Panel A). Although the oligomer appears loosely associated, the large surface area (1,240 Å²) buried at the subunit interface supports this being the native arrangement for the lectin. Corroboration by SEC (fig. 80, Panel B) of the molecular mass estimated for the crystallized structure (51 kDa) also supports the accuracy of this subunit arrangement. Surprisingly, this model contradicts the conclusion others have previously reached based on electrophoresis results notwithstanding similar results from SEC [168]. The quaternary structure of the eel agglutinin was concluded to be of a homo-tetramer [168, 182, 382, 383] formed by association of two

cystine-linked dimers. Based on the AAA structure, the presence of an intermolecular cystine bond is unlikely since those present are internal except for that on CDR4, which is directly in contact with the saccharide. An explanation for this SDS-PAGE band pattern is that the covalently linked bands are artifacts created during the electrophoresis process. Similar results were observed when AAA was treated with an expired reducing SDS-PAGE sample solution (data not shown). Cystine-forming positions are well conserved among eel fuclectins, however FTL-1, -3 and -7 (GenBank Accession BAB03524, BAB03525 and BAB03529) possess an odd half-Cys as their last residue, which possibly could form an intermolecular bond. Nevertheless, AAA did not reveal a similar residue and so the contribution of this singular cysteine to quaternary structure of those isolectins remains presently unknown. In recognition of the conflicting reports on the quaternary state of AAA there is the possibility that MsaFBP32 forms a trimer despite high confidence in the analytical analysis described in chapter 2. Alignment of the MsaFBP32 peptide sequence to the AAA structure reveals that the chloride-coordinating Lys, which forms the trimer's symmetry axis, is conserved. This is relevant to the results from the cross-linking analysis since BS3 specifically reacts with primary amino moieties (i.e. NH₂-terminal and lysine). Therefore, if MsaFBP32 forms a trimer, at least two of these three proximal lysines could be cross-linked through the bifunctional reagent, forming dimers as the major species. This was not the case, so the ladder observed by SDS-PAGE must result from heterologous residues present on opposite sides of the monomer, which linked as they collided. A similar analysis has not been performed on AAA to test if it behaves similarly, but the crystallographic structure of MsaFBP32 most likely will provide the most accurate answer to this question.

An interesting revelation by the AAA quaternary structure is its similarity to the symmetry present in lectins involved in surface recognition [51]. This association creates a planar array of binding sites at one end of the oligomer, which has consequences on surface selectivity and affinity. For example, the spacing between the binding sites of MBPs (45-53 Å) is broader than between residues of an oligosaccharide on a glycoprotein [384] so it minimizes binding to self-glyconjugates. Instead, the wide spacing appears to favor interaction with the arrays of saccharides typical of a pathogen's surface. However, the binding-sites on AAA are much closer (25 Å) suggesting that its target is most likely densely fucosylated. The clustering of protomers also has the effect of compensating [385] for the weak affinity (mM) typically exhibited by individual binding sites of lectins. Association in higher order multimers potentially multiplies individual site affinities. The mechanism of agglutination (i.e. bridging) by AAA of RBCs remains unexplained by the redefined quaternary structure since binding sites are not opposed so as to link cells. Maybe the coating of the cells with lectin diminishes the repulsion between the charged surfaces. However, the internal dihedral point symmetry of the tandem domains in MsaFBP32 would produce such a bridge if it were to associate as a trimer. Clearly, AAA exhibits many of the same topological properties as other lectin families, which contribute to efficient binding.

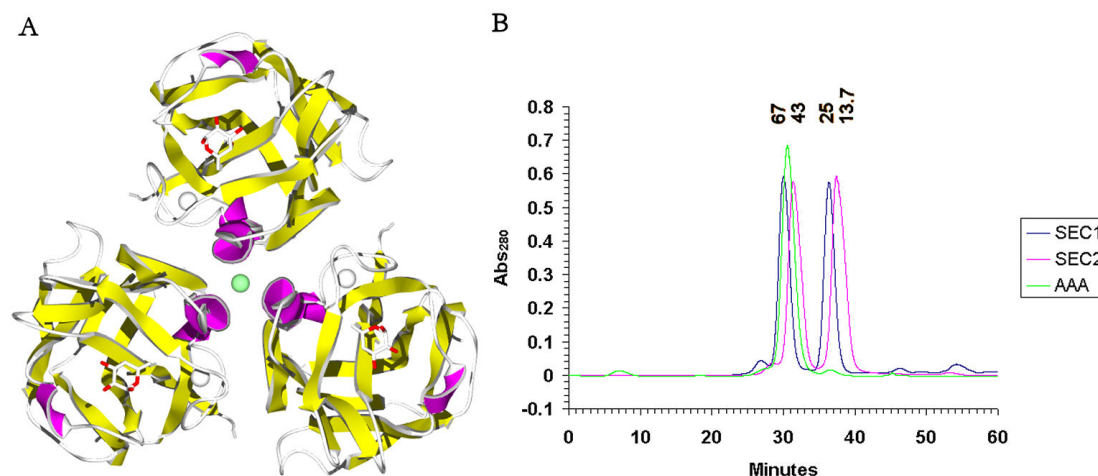


Figure 80. Quaternary structure of AAA.

A. Ribbon diagram with helices colored magenta and β -strands colored yellow. The single chlorine ion marking the three-fold axis of rotation is coordinated by a lysine (Lys₁₆) from each subunit. Bound fucose is shown as a stick model. A white sphere illustrates the calcium. **B.** Analysis of AAA size by SEC. The lectin peak (green) elutes between that of BSA (67kDa) and ovalbumin (43 kDa) and consequently its estimated size is 53.3 kDa, representing the size of a trimer.

A structural view of the FBPL motif

As discussed, the elucidation of the AAA 3-D structure provides valuable details explaining the basis of its biochemical properties and, possibly biological function(s). Consequently, evolutionarily conserved variations in the FBPL motif can be analyzed in detail focusing on positions relevant for ligand binding, metal binding, or tertiary structure (fig. 81) and serve as a guide for future experimentation. Evident from the alignment is an α -Fuc-binding motif (HX24RXDX4(R/K)) shared by most domains including those from proteins present in prokaryotes, which suggests that most FBPL domains are specific for this saccharide.

However, some domains deviate from this motif, suggesting alternate specificities or even lack of saccharide-binding activity. An example is the fruit fly's CG9095 where two residues of the sugar-binding triad are mutated to aliphatic residues unable to produce the H-bonds typical of the AAA binding site. This reduces the likelihood that the domain binds sugar, coinciding with the observation previously discussed that CG9095's C-lectin domain also is unlikely to

bind carbohydrate. The case of both domains of the binary DreV appears to be extreme since a large deletion is present in the CDR4 loop, which would most likely abrogate α -Fuc-binding. Another interesting pattern is that most binary FBPLs (i.e. striped bass, zebrafish, pufferfish and stickleback) possess a unique combination of saccharide-binding motif and cystine bonds (NH₂-term: CXHX24RGDCCXERXX16XX22C) and (COOH-term: CXHX24RDXXXERCX16CX22C). Specifically, one domain has lost the contiguous cystine that contacts the saccharide ring while it gained the nested cystine. It is not clear what effect the loss of this cystine has on binding since the alkaline residue trio is conserved but the predominance of this character only among binary FBPLs suggests that it has some functional value. Future recombinant expression of individual domains should allow biochemical confirmation of these possible alternate specificities.

Analysis of the mutation effects on metal-binding residues is not as clear as for saccharides since most ligands are provided by main-chain carboxyls and their conservation probably relates more to their configuration in the folded protein. However, differing degrees of conservation are evident for the three positions that interact through side-chains. Serine 49 is of special interest, since as a bidentate ligand, it is central to the pentagonal bipyramidal coordination geometry. In a few cases, this position is substituted with non-alcohol residues unable to produce the bidentate coordination (i.e. His, Asp, Pro, Gly, and Val). In most others however, it is substituted by residues possessing oxygens (i.e. Asp, Gln, Glu, Thr and Tyr) that likely are still able to form coordination bonds. In the former cases, there is the possibility that a water molecule may substitute in coordination or there is a reduction in coordination. A coordination number of six produces an octahedral geometry, which for Ca²⁺ is permissible as exemplified by several solved protein structures [381]. Nevertheless, there appears to be a

strong selection for heptacoordination in FBPLs to stabilize the large helix-rich loop. Similar support is given by the absolute conservation seen for Glu 147, which also provides a side-chain ligand.

As commonly observed in other proteins, sequence insertions or deletions (indels) are most permissible in loops where they have minimal effect on the core fold. Such is the case for the F-type lectins where the CDRs are the sites for various indels. Interestingly, the CDR1 loop, which appears to interact with saccharides subterminal to α -Fuc, shows great divergence suggesting that it might regulate binding to the wide diversity of glycoconjugates.

Coincidentally, the intron that splits the exons coding for each domain of MsaFBP32 and others is localized (AAA: Glu 123-Cys 124) close to the long irregular turn at the lower side of the barrel which also presents length variability but it does not appear to be subject to junctional diversity during splicing.

Figure 81. Conservation of functional positions within the FBPL family.

Sugar-binding residues are highlighted in yellow. Calcium-binding residues are highlighted in blue. Cysteines are colored in red. Residues that interact with the bound metal through either the backbone oxygen (B), side chain oxygen (S) or both (B/S) are highlighted in blue below the alignment. Half-cysteines are highlighted in red and italic numbers below the alignment indicates partners. Residues involved in hydrogen bonding to Fuc are highlighted in yellow. Secondary structure of AAA is illustrated above the alignment. Spirals represent α helices and arrows represent β strands. Sequence numbering starts from the first amino acid of the NH₂-sequenced protein or predicted cleavage site of the deduced full primary sequence. Abbreviations: Xla, *X. laevis*; Xtr, *X. tropicalis*; Msa, striped bass; Mch, white bass; Gac, stickleback; Omy, steelhead trout; Dme, fruit fly; Aga, tiger mosquito; Dre, zebrafish; Cca, common carp; Fru, tiger pufferfish; Tni, spotted green pufferfish; Aja, Japanese eel; Ttr, Asian horseshoe crab; Dja, planarian; Cvi, oyster; Mde, *Microbulbifer degradans*; Spn, *Streptococcus pneumoniae*. Alignment was produced with Clustal_X v.1.81 [203] and shaded with GeneDoc v.2.6.002 [204]. Consensus is illustrated on bottom row by lowercase letters for the most frequent residue and numerals indicating Blosum 62 matrix similarity groups (i.e. 6: LIVNM).

Structural analogy is widespread

The lack of detectable similarity of F-type lectin motif to other protein families would suggest that this fold is unique. To test this hypothesis, a structure-based comparison was made by means of the Dali database [386]. Surprisingly, at least seven other proteins with negligible sequence similarity shared the same jellyroll fold as the eel fucoselectin (fig. 82). The structural analogs are the C domains of human blood coagulation factor V (FA58C) and VIII (PDB 1CZT) [387], the COOH-terminal domain of a bacterial sialidase (PDB 1EUT) [281], the NH2-terminal domain of a fungal galactose oxidase (PDB 1GOF) [282, 388], a subunit of the human APC10/DOC1 ubiquitin ligase (PDB 1XNA) [389], and the NH2-terminal domain of the XRCC1 single-strand DNA repair complex (PDB 1JHJ) [390]. Curiously, the FA58C domain of identical structure is present next to the FBPL domain in the *M. bulbifer* ORF previously described. Evident from the sequence alignment derived from these superimposed structures is that although they only share marginal sequence similarity among themselves (fig. 83, Panel A) they are still able to fold into virtually identical domains. Clearly, the fold is better defined by a physicochemical profile rather than by the conventional similarity matrix typically used for determining homology (fig. 83, Panel B).

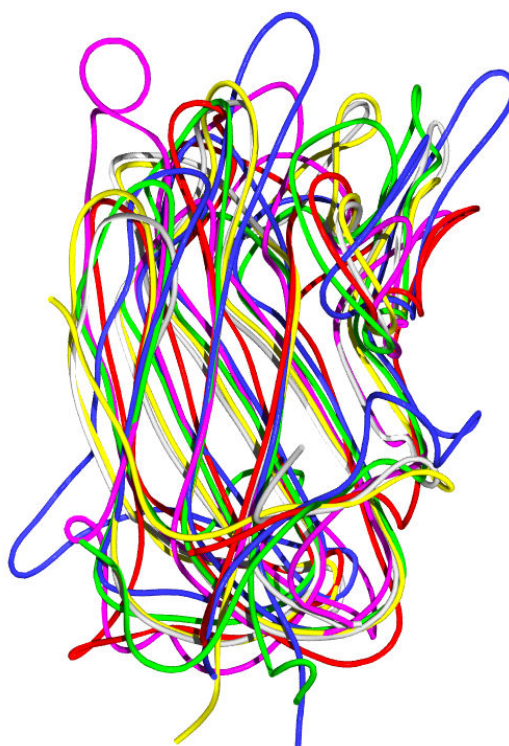


Figure 82. AAA shares structural similarity with diverse proteins.

Green, AAA (PDB 1K12); Yellow, fungal galactose oxidase (PDB 1GOF); Gray, bacterial sialidase (PDB 1EUT); Blue, C2 domain of coagulation factor V (PDB 1CZT); Red, anaphase-promoting complex (PDB 1JHJ); Magenta, DNA single-strand break repair complex (PDB 1XNA).

Five distinct protein families (i.e. galactose binding, discoidin, AAA, XRCC1, and APC10/DOC) are represented by the structural alignment with AAA; nevertheless, their homology is questionable. Surveys of sequenced genomes indicate that frequently protein families share common folds (superfolds) but that most folds are unique to a protein family [391]. As of the latest release of SCOP (v1.63) [392] a total 15 families including AAA have been assigned to the galactose-binding domain-like fold. However, most of these structures are likely the result of convergence resulting from the limited number of favorable topological folds [393] since they only share the 5/3 β -sheet with AAA but not the exact jellyroll topology. The strongest evidence of ancestry comes from shared saccharide-binding residues in members of the galactose-binding family. The secreted bacterial sialidase and fungal galactose

oxidase possess similar enzymatic domains linked to the galactose-binding domain, but their positions on the polypeptide are different, which suggests they have emerged independently [394]. The lectin domain apparently serves to enhance the interaction of the catalytic domain with the substrate similar to the previously mentioned family 6 CBM from *C. stercorarium*, which coincidentally also shares a jellyroll topology [395]. However, the interaction between the Gal crystallized with bacterial sialidase does not explain binding specificity since only the equatorial hydroxyl on C-3, commonly shared Glc and Man [50]; no saccharide heteroatom is evident in the galactose oxidase structure to facilitate corroboration of these findings.

Other structural analogues may have originated from carbohydrate-binding domains but have evolved new specificities. It also appears that all of these families share the placement of their binding-site among the loops at the opening of the β -barrel. The FA58C domain, present as a tandem repeat at the C-terminal of part of multi-domain circulating coagulation factors, is similar to the discoidin domain (DS) [396] named for a lectin from the slime mold [397, 398] for which the binding site has not been determined. Interestingly, like the F-type domain, the FA58C domain is also frequently found as tandem replicates. However, instead of binding saccharides, its CDRs are studded with hydrophobic residues that enable it to bind phospholipids present on the cell surface [399]. In contrast, the b1 domain of neuropilin-1 receptor [400], a DS homologue, is devoid of the hydrophobic spikes and instead presents a polar cleft. Another set of DS homologues, the discoidin domain receptors [401], which possess tyrosine kinase-signalling domains, are activated by binding to collagen [402]. It is evident from these examples that the animal DS domains have evolved to bind a diversity of ligands other than saccharides. The remaining structural analogs, among the only jellyroll proteins placed cytoplasmically [389], do not appear either to bind glycoconjugates. For

XRCC1, the binding site for single-strand DNA breaks was mapped to a groove at the top of the barrel, but does not appear to involve interactions with the glycoside backbone. No binding activity has been attributed to APC10/DOC1 domain, however the variable loops suggest it has a putative ligand. Curiously, the presence of a cation-binding loop is variable among these structural analogs since FA58C, APC10/DOC1, nor XRCC1 have a cation present in their elucidated structures.

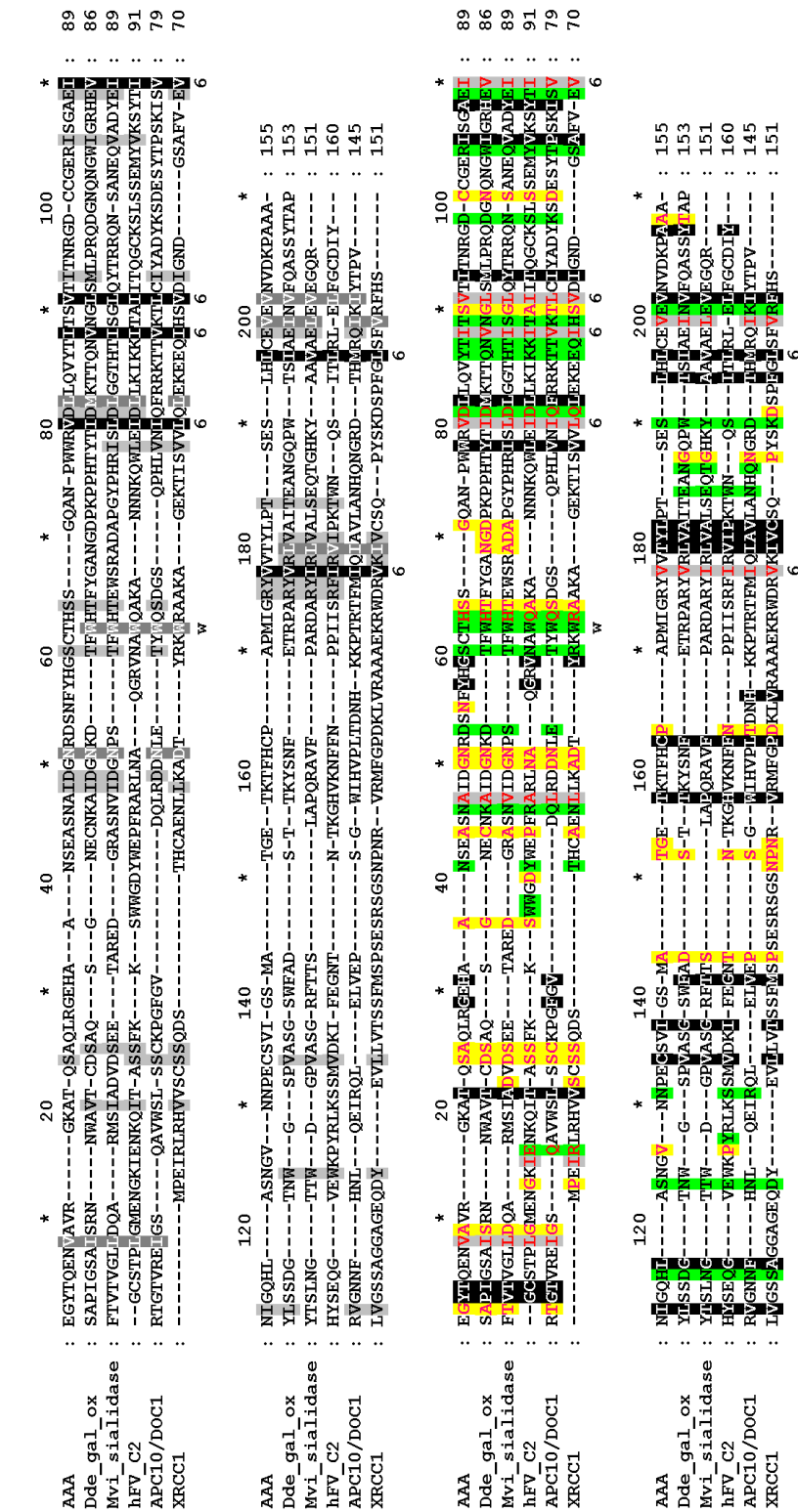


Figure 83. Structure-based sequence alignment of AAA and its structural analogs.

A. Upper alignment illustrates similarity (Blosom62). Positions are shaded according to degree of conservation among sequences: in 100% of sequences are shaded in black, in ≥ 80 of sequences are shaded in gray with white lettering and in $\geq 60\%$ of sequences are shaded in gray with black lettering. Sequence numbering starts from the first amino acid of the N-sequenced protein or predicted cleavage site of the deduced full primary sequence. Sequence names: AAA, European eel agglutinin (PDB 1K12); Dde_gal-ox, fungal galactose oxidase (PDB 1GOF); Mvi_sialidase, bacterial sialidase (PDB 1EUT); hfv_C2, C2 domain of coagulation factor V (PDB 1CZT); APC10/DOC1, anaphase promoting complex (PDB 1JHJ); XRCC1, DNA single-strand break repair complex (PDB 1XNA).

B. Lower alignment illustrates conservation of physicochemical properties: Small (red letters/yellow background), Aliphatic (red letters/grey background), Hydrophobic (white letters/black).

In summary, resolving the AAA 3-D structure has substantially added to the understanding of both biochemical features and ancestry of FBPLS. Clearly, the motif represents a structural domain unique among all other known animal lectin families. This F-type lectin fold exhibits a compact jellyroll topology formed by two sandwiched β -sheets and is stabilized by cystines. Selectivity for the specific features of α -Fuc conformation is achieved mostly through hydrogen bonds provided by a highly conserved trio of basic residues, which center on the axial hydroxyl of C-4. Further selectivity for fucosylated complex glycoconjugates is expected from sterically inevitable interactions with the extended variable loops surrounding the binding site and a unique cystine formed by contiguous residues. Calcium is present in the lectin, but unlike the C-type domain, not as a saccharide-coordinating ligand but as a structural prosthetic group. Mutability of the identified saccharide-binding motif in members of the family suggests that ligands other than to α -Fuc are likely. Although the fold is unique among animal lectins, surprisingly four other widely distributed protein families with diverse functions share it. Microbial saccharide hydrolases still appear to apply the domain for binding carbohydrates. However, in metazoans these analogs apparently have diverged to bind anionic surface lipids or intracellular ligands such as DNA. Clearly, in contrast to the FBPLs, the fold is present in higher vertebrates but the ancestry among the families is debatable since sequence similarities are negligible. It is certain that this fold emerged early in life yet may have been adopted independently by each of these distinct protein families. The current high priority placed on elucidation of representative protein 3-D structures should quickly provide a complete assessment of the diversity families adopting this fold.

CONCLUSIONS AND FUTURE DIRECTIONS

As planned, the approach of using affinity chromatography for screening for saccharide-binding proteins from blood or liver of the bass was effective in pointing to the presence of a fucose-binding lectin in circulation. Biochemical characterization (i.e. SDS-PAGE, SEC, ESI-MS and cross-linking) of this lectin indicates that it consists of a 32 kDa monomer (FBP32). Although similar in molecular size to the typical size expected for the reduced MBL monomer, FBP32 does not form a covalently-linked trimer or appear to assemble into higher order oligomers characteristic of collectins. Cloning of the gene through PCR using peptide sequence-based primers and liver cDNA provided the initial gene sequence, which subsequently allowed deduction of the full polypeptide sequence. Evident in the ORF is a tandemly duplicated motif (FBPL) that does not share sequence similarity to C-type lectins or any other lectin family motifs presently described. Initial searches in the gene repositories for proteins homologous to FBP32 identified only one cDNA (XL-PXN1) isolated from the liver of the African clawed frog [209]. Interestingly, the sole FBPL motif present in this frog homologue was adjacent to a pentraxin domain. The human CRP, which is a homomultimer of the pentraxin domain, is well described as positive acute phase protein meaning that it is greatly upregulated during inflammation. Recent investigation links chronic inflammation to an increased risks of developing heart disease, so CRP has become a valuable measure for the prognosis of heart infarction [403]. Moreover, mammalian pentraxins are capable of pathogen recognition, mediated through lectin activity that leads to complement cascade fixation. Reports of polypeptides similar to FBP32, but consisting of a single FBPL, from the eel and the horseshoe crab provide evidence linking these lectins to pathogen recognition [181, 182].

This supports a role for this unique family of lectins as humoral factors of the innate immune system whose expression is potentially upregulated during inflammation. Preliminary study of the effect of inflammatory challenge in the bass point to an increase in gene transcription, but no obvious increase of lectin levels in circulation was evident. Current understanding of inflammation comes from studies in mice and humans but little is known of inflammatory processes in lower vertebrates. Several cytokine homologues have been identified from teleosts that are known to mediate inflammation in mammals [404], but a readily detectable marker of inflammation, such as CRP in humans, has not been established for teleosts. Development of such a sensitive marker would facilitate determining how levels of FBPLs correlate to a developing inflammation. Pentraxins, ficolin and collectins in addition to activating complement also serve as opsonins [405-407]. The receptors that have been implicated in recognition of particles coated with these proteins are the Fc receptor and the C1 receptor [408]. There is no structural similarity apparent between the Fc region of immunoglobulins and pentraxins, yet opsonophagocytosis of pentraxins is convincingly mediated by Fc receptor. Therefore, it is plausible that MsaFBP32 may also share a receptor with other opsonins. Phagocytosis assays may provide an indication of this activity. Clearly, further study of FBP32 as a potential opsonin is needed to unequivocally assign its function to innate immunity.

The four diverse FBP homologues initially identified hinted that the domain could be shuffled as a unit to create diverse configurations. To gauge the diversity of FBPLs and attempt to gain functional information through phylogenetic comparison, incomplete sequences (i.e. ESTs) showing similarity to the FBP domain were completed to deduce the full polypeptide sequence. The rapid advance of genome sequencing also provided the opportunity for

enumerating the full set of FBPL genes present within a species. Additional configurations were positively identified including tandem repeats of two to five FBPLs and a likely cell surface receptor with additional C-type lectin and sushi domains. From this compilation, it became evident that FBPLs are present in select eubacteria, in both lophotrochozoan (i.e. mollusca and platyhelminthes) and ecdysozoan protostomes (i.e. merostomata and insecta), early deuterostomes (i.e. echinodermata) and the already mentioned lobe-finned fish lineage (i.e. *X. laevis*) and ray-finned fishes. In contrast, as confidently supported by whole genomes sequenced in some cases for several congeners, this lectin family appears absent from fungi, nematodes, ascidians, and higher vertebrates (i.e. mammals and chicken) which suggest that it has been selectively lost independently several times even in relatively closely-related lineages. The paucity of bacteria possessing an FBPL also suggests that either many lineages also lost this domain, which is a parsimonious scenario in this kingdom, or that it may have been acquired through horizontal transfer from metazoans. Lack of shared FBPL domain configuration among the phyla sampled makes tracing ancestry difficult for this lectin family except in euteleosts where the binary domain configuration is shared. Nevertheless, even these FBPLs point to lineage-specific duplications within the diverse teleost orders represented so inter-taxon correspondence of genes is still not evident throughout the tree. The mottled phylogenetic distribution of this lectin family, diverse temporospatial expression, and its varied domain architecture point to a functionally plastic protein, which has been tailored in each lineage and apparently sometimes has lost its fitness value. The absence of FBPLs from higher vertebrates is an evolutionary quandary that eventually could be resolved through targeted sampling of taxons neighboring the Anura but the great extinction events of the past that claimed so many families of amphibians may have obscured any potential evidence forever.

With the accrual of genome sequences, further insight is possible into the evolution of FBPL genes and their transcriptional control. The sequence of the duplicated striped bass FBPL genes illustrate the common exon/intron organization shared in this lectin family in addition to evidence of a localized gene expansion, a feature not present in the pufferfish FBPL loci. Each lectin domain is encoded by two exons and flanked by phase 1 introns, a characteristic apparently conducive to exon shuffling since it is shared with domains frequently present in tandem configurations within numerous blood proteins. Regarding control of gene expression, computational analysis of the regions proximal to the putative transcription start sites detected an Inr motif in MsaFBP32 only, but no other motifs common to core promoters or involved in inflammation (i.e. NF- κ B). Both striped bass genes described are coexpressed in the liver so they likely share similar control elements. Surely, the proximity between the striped bass genes should facilitate promoter mapping, but presently no immortalized cell cultures have been developed from the liver of this species, which hinders this approach. Long-distance alignment of the pufferfish orthologues detected conserved stretches of upstream non-coding DNA that may be involved in control of gene transcription. The release of the zebrafish genome is drawing near, which should allow testing if these motifs are conserved in a more distant orthologue, therefore providing stronger evidence of their positive selection. Considering the multiplicity of FBPLs and absence from higher vertebrates, thorough analysis of *in vivo* function would benefit greatly from development of gene “knockout” methods for lower vertebrates. Developmental patterns of expression of FBPLs are presently unknown but “knockdown” methods, currently applicable only during early development, are an alternative approach for FBPLs that are developmentally regulated.

Recent developments in non-lethal screening for effects of knockdowns in adult mice [409]

holds promise for studying FBPL function in adult fish.

Resolution of the crystallographic structure of the eel agglutinin, a single-domain FBPL, greatly advances understanding this lectin family as it finally translates from motif into a unique spatial topology designated F-type fold. In addition, co-crystallization of the lectin with α -Fuc provides structural evidence explaining the specificity exhibited by this lectin. Specifically, a trio of well-conserved basic residues in a shallow pocket and a unique contiguous-residue cystine directly interact with the uniquely configured natural saccharide. However, other interactions may also form with unrelated ligands, which mimic some of the characteristics of α -Fuc expanding the range of possible ligands. The variability in the loops surrounding the binding pocket also suggest they could exert selectivity on ligand repertoire in analogy to the CDR loops of Igsf. Another novel discovery obtained from the crystal structure is that the lectin associates into a previously unknown quaternary structure of three-fold cyclic symmetry very similar to that observed in collectins. Surprisingly, topological comparison indicates that the fold is more widespread than indicated by sequence homology alone. Structural analogs include vertebrate coagulation factors, eukaryotic DNA repair enzymes and cell cycle control proteins, and microbial hydrolytic enzymes demonstrating the emerging trend of topological redundancy between sequence-based families. Presently, the crystallographic structure of MsFBP32 has also been resolved and illustrates the binding-sites of the tandem domains as being antipodal. It also confidently demonstrates MsFBP32 as forming a trimeric quaternary arrangement like that observed for AAA, which disagrees with the results from SEC.

Evidently, readily applicable biochemical methods are insufficient to accurately describe the physicochemical properties of FBPLs. Comparison of FBPL homologues to the AAA

structure suggests most of them probably bind α -Fuc, but a few may have developed alternate specificities. With the accumulated evidence in hand, it is now possible to address the contextual diversity of FBPLs with a view to identify their fucosylated ligands more effectively. The findings presented here open several distinct avenues for future studies. The moronid bass are advantageous as accessible and copious sources for blood, but due to their lengthy reproductive cycle are not suitable as genetic models. Several other teleosts have been proposed as suitable biological models but have generated limited interest in the general scientific community. Currently, advances in biomedicine using zebrafish as *in vivo* models of vertebrate early development have broadened the tools available for studying gene function, but *in vitro* assays for studying cellular function still require further development. In comparison to mammalian cell lines very few teleostean cell lines have been established and been made generally accessible. Currently, immortal cell lines from trout liver (SOB-15) and zebrafish (ZFL) are available, but do not appear to express FBPL under normal culture conditions (data not shown). Establishing conditions that induce FBPL expression in these cells would provide an important step for tracing the signaling network that regulates these genes and eventually correlating it with the physiological state of the organism. Structural studies are greatly advanced with the two unique FBPLs crystallized but still many questions remain unanswered. Formation of the trimer that is apparent for AAA by SEC but not for MsaFBP32 is still perplexing. The ability of analytical ultracentrifugation to allow direct observation of sample aggregation states under equilibrium is the best candidate approach to address this question. The role in binding of the contiguous-residue cystine present in some FBPLs, but not others should be further addressed to determine its contribution to ligand affinity. This raises the issue that the kinetics and energetics of binding have yet to be

addressed. Kinetic analysis of ligand binding using optical biosensors has become commonplace nowadays as the technology can accommodate diverse ligand-receptor species and has been applied to lectins. Titration microcalorimetry should also provide useful information regarding the driving forces behind the observed binding selectivity. Currently, attempts are being made to purify recombinant FBPLs from a bacterial host, but a refolding step will likely be required to obtain active lectin. This is important given the fact that FBPLs come from smaller organisms that require special keeping. Finally, the vagaries in inheritance of FBPLs should serve as a warning. If conceited evolutionary suppositions aim our scientific gaze only onto ourselves much biological knowledge could be missed.

APPENDICES

A. MsaFBP32 cDNA

```

      ATGTCTCAAAGAGGCTAGGTACAGTAACTCACTGGACTCCAGGGATAAAAGATCTGTTCTAACCAGGAAG
1  -----+-----+-----+-----+-----+-----+-----+-----+
      CAGGAATAATGAGGCACAGTGTGGTATTTCTGTTGCTGCTCCTCTTAGGGGCGTGTTCAGCTTACAACCTA
71 -----+-----+-----+-----+-----+-----+-----+-----+
          m r h s v v f l l l l l l g a c s a Y N Y

      TAAAAATGTGGCCTTGCGTGGAAGCGACTCAGTCGGCACGTTATTTGCACACACATGGAGCCGCCTAC
141 -----+-----+-----+-----+-----+-----+-----+-----+
          K N V A L R G K A T Q S A R Y L H T H G A A Y

      AACGCCATTGATGGAACCGTAACTCTGACTTCGAAGCTGGATCGTGCACCCACACTATTGAACAGACCA
211 -----+-----+-----+-----+-----+-----+-----+-----+
          N A I D G N R N S D F E A G S C T H T I E Q T N

      ACCCCTGGTGGAGAGTGGACCTACTGGAGCCCTACATCGTCACCTCCATCACCATCACCAACAGAGGAGA
281 -----+-----+-----+-----+-----+-----+-----+-----+
          P W W R V D L L E P Y I V T S I T I T N R G D

      CTGCTGTCCAGAAAGGCTCAACGGGGTGGAGATTACATCGGCAACTCTATACAAGAAAATGGTGTGCA
351 -----+-----+-----+-----+-----+-----+-----+-----+
          C C P E R L N G V E I H I G N S I Q E N G V A

      AACCCAAGGGTTGGTGTAAATTTCTCATATCCCTGCAGGGATCTCACATACTATCAGTTTCACTGAACGTG
421 -----+-----+-----+-----+-----+-----+-----+-----+
          N P R V G V I S H I P A G I S H T I S F T E R V

      TGGAGGGACGTTACGTGACTGTGCTTCTACCTGGTACAAACAAGGTTCTTACACTCTGTGAAGTGGAGGT
491 -----+-----+-----+-----+-----+-----+-----+-----+
          E G R Y V T V L L P G T N K V L T L C E V E V

      TCATGGGTACCGAGCCCCAACTGGAGAGAACCTGGCCCTCCGAGGAAAAGCCACACAGTCTTCATTGTTT
561 -----+-----+-----+-----+-----+-----+-----+-----+
          H G Y R A P T G E N L A L R G K A T Q S S L F

      GAATCTGGTATTGCATATAATGCCATTGATGGGAATCAAGCCAACAATTGGGAAATGGCCTCCTGCACTC
631 -----+-----+-----+-----+-----+-----+-----+-----+
          E S G I A Y N A I D G N Q A N N W E M A S C T H

      ACACAAAAAACACAATGAACCCCTGGTGGCGAATGGATCTGAGCAAAACCCACAGAGTGTTCCTGTAA
701 -----+-----+-----+-----+-----+-----+-----+-----+
          T K N T M N P W W R M D L S K T H R V F S V K

```

```

771  GGTAACCAACCGAGATTCATTTGAAAAACGAATCAATGGAGCTGAGATCCGAATTGGAGATTCCCTCGAC
    -----+-----+-----+-----+-----+-----+-----+
      V  T  N  R  D  S  F  E  K  R  I  N  G  A  E  I  R  I  G  D  S  L  D

841  AACAAACGGCAACAACAATCCCAGGTGTGCTGTGATCACAAGCATCCCAGCAGGTGCTTCTACTGAATTCC
    -----+-----+-----+-----+-----+-----+-----+
      N  N  G  N  N  N  P  R  C  A  V  I  T  S  I  P  A  G  A  S  T  E  F  Q

911  AGTGTAACGGGATGGATGGCCGCTATGTTAACATTGTTATCCCTGGAAGAGAAGAGTACCTGACCCTGTG
    -----+-----+-----+-----+-----+-----+-----+
      C  N  G  M  D  G  R  Y  V  N  I  V  I  P  G  R  E  E  Y  L  T  L  C

981  TGAGGTGGAGGTGTATGGCTCTGTCCTGGATTAGGTGTCAGTACTAATACTGTTGAATGTACACAAACAA
    -----+-----+-----+-----+-----+-----+-----+
      E  V  E  V  Y  G  S  V  L  D  *

1051 AACAAAATAGTAGATTAAGCTTTTTTGATTGTTTCCATTCAAATAAGACAGAGATGGTCTTATCCAATA
    -----+-----+-----+-----+-----+-----+-----+

1121 AAATTACATCACGAAAAAAAAAAAAAAAAA
    -----+-----+-----+-----+-----+-----+ 1149

```


B. Xla-PXN1 cDNA

```

      10              30              50              70
      .              .              .              .
AGAGCTACTGGGACATCCAGTCTATGCAACAGAGCTCTCCAGTATGGCATCTTCCATATTTTGTATACTC
      .              .              .              .
      M A S S I F C I L
      90              110              130
      .              .              .              .
TTGTATTTAGGCATTATTTGTGGACAACACATTGGAGTTGATGCCTGGAGTCCTCCTAAAGATGGACCCT
L Y L G I I C G Q H I G V D A W S P P K D G P Y
      150              170              190              210
      .              .              .              .
ATTTTAAGCCAGCCCCTAATGTGGCACTTGATGGAATTACTTCACAGTCCAGTACCATGGCCTATTATGG
F K P A P N V A L D G I T S Q S T M A Y Y G
      230              250              270
      .              .              .              .
AAACTCAAGACATGCCAACGATGGTTCTTTGGCCAACAACCTATCTGAGATCTCAGTGCTCCTACACAAA
N S R H A N D G S L A N N Y L R S Q C S Y T K
      290              310              330              350
      .              .              .              .
AAAGAAGCAGACCCCTTGGTGGATGGTGGACCTACAGAAACCTTATCAAATTTTATCTGTGGCTGTCACCA
K E A D P W W M V D L Q K P Y Q I L S V A V T N
      370              390              410
      .              .              .              .
ACAGAGTGTTGGAATGCTGCAAAGAAAGGCTTTTTAATGCTGAAATCCACATTGGAAATGACCCTAAGCA
R V L E C C K E R L F N A E I H I G N D P K Q
      430              450              470              490
      .              .              .              .
AGGTGGAATAATTAATCCCAGATGCGGTGTTATCTCATCTATAGAATCTGGGGAGACCCCTTTCATTCTCA
G G K L N P R C G V I S S I E S G E T L S F S
      510              530              550
      .              .              .              .
TGCCAAGGAATGGTTGGGCAGTATGTGACCATTACTTTACCAGGGAAGGAGGAGCATCTTATTCTGTGTG
C Q G M V G Q Y V T I T L P G K E E H L I L C E
      570              590              610              630
      .              .              .              .
AAGTTC AAGTGT TGGTCTGCCTGT CAGCAGT TCTGATGAT GTTGAAGTGACGGCACCAAAATATCTGAC
V Q V F G L P V S S S D D V E V T A P K Y L T
      650              670              690
      .              .              .              .
AACACCAAACGAGCTCCAAACTTGGCTGTGAAAGGGATAGCCCAGCAATCCAGTCTCTACAACATGTAT
T P N G A P N L A V K G I A Q Q S S L Y N M Y
      710              730              750              770
      .              .              .              .
GGAGAACCAAAGAACGCCAATGATGGGTCTCTAGCCAGTAATTATTTCTTCTTGAGTGTGCCAGCACTA
G E P K N A N D G S L A S N Y F F L E C A S T S
      790              810              830
      .              .              .              .
GTGAGCAGGAGGATCCCTGGTGGATGGTTGATCTCAAAGCAAGCCACAGAGTTTACACTGTAGCTGTGAC
E Q E D P W W M V D L K A S H R V Y T V A V T
      850              870              890              910
      .              .              .              .
CAACAGAGGTGACTGCTGTGCTGAGAAAATTAACAATGCTGAAATAAGAATCGGAGATTCCAACGATGCA
N R G D C C A E K I N N A E I R I G D S N D A
      930              950              970
      .              .              .              .
GGAGGACAACAAAATCCAGTATGTGGCATTATCAAGTCAATGGCCAATGGGGAGACACTTTCCTTTGAGT
G G Q Q N P V C G I I K S M A N G E T L S F E C
      990              1010              1030              1050

```

GCAATGGCATGCAAGGTCAGTATGTGACTGTCTTTATCCCTGGAAATAAAACATCACTCACCATCTGTGA
 N G M Q G Q Y V T V F I P G N K T S L T I C E
 1070 1090 1110

AGTGCAAGTGTTTGGCCTCTCTAGTGAAGCTCCTGATTATACTGGAATATATGTGGTTTCTAAAGATGAC
 V Q V F G L S S E A P D Y T G I Y V V S K D D
 1130 1150 1170 1190

TCATTTACCTGGCCGACATATTTATAAATTTTTTGGTCTATGGAGCAGTGACCAAGAATACGATTATG
 S F H L A D I F I N F F G L W S S D Q E Y D Y D
 1210 1230 1250

ATTTACCTACAGTAGCTACACGAACTGATGAAAATCTTGCCTTCAGGGGAATCAGCTCCCAGTCCAGCAC
 L P T V A T R T D E N L A F R G I S S Q S S T
 1270 1290 1310 1330

ATATGATAACCTTGGAAGGCAGAAAATGCCATTGATGGTTCCACCAGTACAAAATATATGTCAACGCAC
 Y D N L G K A E N A I D G S T S T K Y M S T H
 1350 1370 1390

TGTTCCTCACACTGACCTAGACATAGAGCCATGGTGGAAAGGTGGACCTTATCAATACCTATAATGTCACCG
 C S H T D L D I E P W W K V D L I N T Y N V T E
 1410 1430 1450 1470

AAGTACAAATAACTAACAGGGGCGACTGCTGTAACAATCGTATTAAATGGTGCAGAAATTAGAATAGGCAC
 V Q I T N R G D C C N N R I N G A E I R I G T
 1490 1510 1530

TGCTCCTGAAAAGGGTGGAAACAAAAATCCCAGATGTGCTAAAATTGCAACCATGGCCCTTGGAGAGTCA
 A P E K G G T K N P R C A K I A T M A L G E S
 1550 1570 1590 1610

GCAACATTTCAGCTGTGGAATGGTGGGGCGATATGTGACTGTTACAATACCAGGCAGAGCTGCCTATCTCA
 A T F S C G M V G R Y V T V T I P G R A A Y L T
 1630 1650 1670

CTCTCTGTGAAGTAAAGGCCTTCGGCCATGAAATTTCTGGAAACTACACTAATAATCCCAGTTCTCCTGA
 L C E V K A F G H E I S G N Y T N N P S S P D
 1690 1710 1730 1750

TTCGAGAGAAATTGAAGAACAGCAAGCTGCCACTGAACTGAGAAATATTTTAAAGCACTCTGATGCCGCA
 S E E I E E Q Q A A T E L R N I L K H S D A A
 1770 1790 1810

TCTAATGTAGCTCTGCATGGGGCTGCATATCAGTCCAGCACAGCTGGTGAAGCTAATGCAAAGAATGCTG
 S N V A L H G A A Y Q S S T A G E A N A K N A V
 1830 1850 1870 1890

TAGATGGCAAGCTACAGAATCAGAATCCAGCTAAACAATGTGCTCAAACAACAGTGGAAACTGATCCCTG
 D G K L Q N Q N P A K Q C A Q T T V E T D P W
 1910 1930 1950

GTGGACAGTTGACTTAACATCAATCCATAAGGTCTTCTCAATTGCGGTGACAAACAGAGGAGACTGCTGC
 W T V D L T S I H K V F S I A V T N R G D C C
 1970 1990 2010 2030

AGTGAAGGGCTTGATGGAGCAGAGATTCACTTAGGAGATTACGCCTTCAGCTGGAAGAAGAACCCTGTGT
 S E G L D G A E I H L G D S A F S W K K N P V C
 2050 2070 2090

GTGGAACCTGTATCCAAAATTGGTCCTGGAGAAACATTTTCTTTTGTGAGTGAATGGAATGGAAGGACGTTA
 G T V S K I G P G E T F S F E C N G M E G R Y
 2110 2130 2150 2170
 TGTAACCATTTGTTCTACTAGGCAATGAAAAGTCTCTTACACTTTGTGAAGTACAGGTCTTTGGCTTAACA
 V T I V L L G N E K S L T L C E V Q V F G L T
 2190 2210 2230
 GTGGAAACACCAAATGGCGAGAGGAATGGTGATTTTGAGCAACAAAAAGAGAATCATGGAGCAAAGAATG
 V E T P N G E R N G D F E Q Q K E N H G A K N V
 2250 2270 2290 2310
 TAGCTCCCCAAGGTATTCCTTATCAGTCAAGCTATTACGGCCAAAAGGAACAAGCCAAACGCGTTATTGA
 A P Q G I P Y Q S S Y Y G Q K E Q A K R V I D
 2330 2350 2370
 TGGATCTCTAGCAAGCAACTACATGGAAGGAGACTGCTGCCACACTGAGAAACAGATGCATCCTTGGTGG
 G S L A S N Y M E G D C C H T E K Q M H P W W
 2390 2410 2430 2450
 CAACTAGACATGAAATCCAAAATGCGTGTACATTCCGTGGCCATCACCAACCGAGGAGACTGCTGCCGGG
 Q L D M K S K M R V H S V A I T N R G D C C R E
 2470 2490 2510
 AAAGAATCAATGGGGCTGAAATCCGCATTGGGAATTCTAAAAAAGAAGGGGGACTTAACAGTACCAGGTG
 R I N G A E I R I G N S K K E G G L N S T R C
 2530 2550 2570 2590
 TGGCGTTGTRTTCAAGATGAATTATGAGGAGACGTTATCCTTTAACTGCAAAGAGCTTGAGGGCCGATAT
 G V V F K M N Y E E T L S F N C K E L E G R Y
 2610 2630 2650
 GTCACTGTAACAATACCAGACAGAATGGAGTACCTAACACTGTGTGAAGTTTCAGGTTTTTGCTGATCCAC
 V T V T I P D R M E Y L T L C E V Q V F A D P L
 2670 2690 2710 2730
 TCGAAGTAGATGGAACGGAAGCATCGGATTCTTCAGAGTCTGTAGATGGAACAGAAGCACCAGCTTCTCC
 E V D G T E A S D S S E S V D G T E A P A S P
 2750 2770 2790
 AGAGTCTGATGTGGAACCTACCAATTGCTTCTGGTATGAATGTGGATTAAACAAACAAATCTTTCATGTTT
 E S D V E L P I A S G M N V D L T N K S F M F
 2810 2830 2850 2870
 CCAAAAAGAAAGTGACATCAACCATGTAAAGCTATTACCTGAAAAAGCAATGAGCCTCAAAGCCTTTACAC
 P K E S D I N H V K L L P E K A M S L K A F T L
 2890 2910 2930
 TCTGCATGAAGGTGCTCTTGAATGTCCCTGAAAATCGAGAAACCATTTTATTTTCATACCGTACAATGTT
 C M K V L L N V P E N R E T I L F S Y R T M F
 2950 2970 2990 3010
 TTACGATGAGCTAAACCTTTGGATTGAGCGTGATGGTAGAATAGGCCTATACATGAGTGGAGATGGCATT
 Y D E L N L W I E R D G R I G L Y M S G D G I
 3030 3050 3070
 ATATTCCTCCGCAATGAAATTTAAAAGTGAATGGAACCATCTCTGTTTGACTTGGGAGTCTAAGTACGGTC
 I F P R M K F K S E W N H L C L T W E S K Y G R
 3090 3110 3130 3150

GCACAGAATTCTGGTTGAATGGCAGACGATCAGCTACAAAAGTGTATCACCAAAGAACACTGTGCGTTC
 T E F W L N G R R S A T K V Y H Q K N T V R S
 3170 3190 3210

AGGTGGAATCGTCTTGTTGGGTCAAGATCAAGATTCTTATGGGGGAGATTTTGATAAGACCCAAAGCTTT
 G G I V L L G Q D Q D S Y G G D F D K T Q S F
 3230 3250 3270 3290

GTGGGACAAATCAAAGATCTTAAGATGTGGAACAAAGTCCTGCCCCTGAGGTCTCTCAAATCTTTGTTCA
 V G Q I K D L K M W N K V L P L R S L K S L F K
 3310 3330 3350

AGGGTAGAGAAATTGGAAATGGCAATATTTTTGATTGGAGTTCATTATCTTACTCAATGATAGGAAATGT
 G R E I G N G N I F D W S S L S Y S M I G N V
 3370 3390 3410 3430

TGCAGAAGTCTAAAATGAAATCAGCTCAAGTTAGTAACACAATGGGCTCTTCTCTGTGCAGTTCTTGAT
 A E V *
 3450 3470 3490

TAGTAACTAACAGTCAAACATCACAGTCCAGAAGCTTACTCTGCGTCTATTCTGTAGAGCTGACAATTA
 3510 3530 3550 3570

AAATAAAATCCATGGTACTCTTGTACGGCTTTGAGTGGCCATATATAGCTTTATATTGTGGTTGATATGT
 3590 3610 3630

TATCACGCTGCATCTTGGGATATATTAGTTAAAGTTAAAAAGATGCTTGTATCTAATTTAGGGAAATAGA
 3650 3670 3690 3710

GAACATGGTCAGTTGTCTGGTCAGATCCTGTGCTTGAAAACGGTTTAGGATTCTGGACTCTGTTTAAACT
 3730 3750 3770

GATTTTACTTTATAACATATCCATTCTGATTGTAAAAATGAAAGTTAAATGTCAAACATAATGTATTAA
 3790 3810 3830

AAAATAAAGATAATACAGTGATTTCAAAAAAAAAAAAAATCGATATCTAGATCTGCGG

C. XlaII-FBPL cDNA

```

      10              30              50              70
      .              .              .              .
GAATGTAGGTTGGCAGAGGACTGCCTGTACCCTACCTGATTGATCCCGATCCCCACACTGGCAACATCTG
      .              .              .              .
      N   V   G   W   Q   R   T   A   C   T   L   P   D   *   S   R   S   P   H   W   Q   H   L
      .              .              .              .
      90              110             130
      .              .              .              .
ACTTGTAAGAATTGATACACCTGTCTACTGATACCTTAAGGATCTGCCTGACTGTGATCATCGCTCTGTC
      .              .              .              .
      T   C   K   N   *   Y   T   C   L   L   I   P   *   G   S   A   *   L   *   S   S   L   C   P
      .              .              .              .
      150             170             190             210
      .              .              .              .
CACCGTGATTCCATCTGCTGTATCCTACAAGTACTTGTCTCCATGTGTCTTAGCTGATCCATAGCAGTTG
      .              .              .              .
      P   *   F   H   L   L   Y   P   T   S   T   C   L   H   V   S   *   L   I   H   S   S   W
      .              .              .              .
      230             250             270
      .              .              .              .
GACAGAAATTGAATGCATCACCCAAGGAAGTGACTATCTCCATAATCTCTGATGATGTGTCTTCTGCTA
      .              .              .              .
      T   E   I   E   C   I   T   Q   G   S   D   Y   L   P   *   S   L   M   M   C   L   L   L
      .              .              .              .
      290             310             330             350
      .              .              .              .
CTCTTGGCTTTTGGAGCCATTGCCCAGGCTCAGAGATGTGATCCCCAGACAGAAGGGGTCAATGTTGCAA
      .              .              .              .
      L   L   A   F   G   A   I   A   Q   A   Q   R   C   D   P   Q   T   E   G   V   N   V   A   R
      .              .              .              .
      370             390             410
      .              .              .              .
GACTGGGAATTGCCAGCCAAAGCTCTACCTATGTCCATGATCCAATGCCTGGTCTGAACGCGCTCTCGA
      .              .              .              .
      L   G   I   A   S   Q   S   S   T   Y   V   H   D   P   M   P   G   P   E   R   A   L   D
      .              .              .              .
      430             450             470             490
      .              .              .              .
TGGAATAACAAGGTCAATGCAATGATCCACCCCTGCTCCACACCTACAATGACTTTGAGCCCTGGTGG
      .              .              .              .
      G   N   N   K   V   N   A   M   I   H   P   C   S   H   T   Y   N   D   F   E   P   W   W
      .              .              .              .
      510             530             550
      .              .              .              .
CGTGTGGACCTGAAAAAGACCTATGCAGTCAACTCTGTGGTCATAGTGAATAGGATGGACTGCTGCAGTG
      .              .              .              .
      R   V   D   L   K   K   T   Y   A   V   N   S   V   V   I   V   N   R   M   D   C   C   S   E
      .              .              .              .
      570             590             610             630
      .              .              .              .
AGCGCCTTGAAGGGGCGCAGGTTTCGTATTGGGAATTCTGCGGACAACAATAACCCAATTTGTGGCACCAT
      .              .              .              .
      R   L   E   G   A   Q   V   R   I   G   N   S   A   D   N   N   N   P   I   C   G   T   I
      .              .              .              .
      650             670             690
      .              .              .              .
CAGTGATGCCTCCCAAGCTACAATCACTCTGTTCTGCAATGGGATGGTGGGTCCGTACCTCAGTGTGTGT
      .              .              .              .
      S   D   A   S   Q   A   T   I   T   L   F   C   N   G   M   V   G   R   Y   L   S   V   V
      .              .              .              .
      710             730             750             770
      .              .              .              .
ATTCAGGACGACAGGAATTTCTCAGCTCTGTGAAGTGGAGGTTTATGGGCAAGAATCTGATGATAAAG
      .              .              .              .
      I   S   G   R   Q   E   F   L   T   L   C   E   V   E   V   Y   G   Q   E   S   D   D   K   D
      .              .              .              .
      790             810             830
      .              .              .              .
ATAATTTGGCGAGACTGGGAGATGCCACTCAGAGCTCAACTTACAGACCCGAGTACAACGCTGGTGCCGC
      .              .              .              .
      N   L   A   R   L   G   D   A   T   Q   S   S   T   Y   R   P   E   Y   N   A   G   A   A
      .              .              .              .
      850             870             890             910
      .              .              .              .
TATTGATGGCAATAAGGTGACAAATATGATGTTGGGCTCATGTTCCCACACCAATAATGACAACCCGGCT
      .              .              .              .
      I   D   G   N   K   V   T   N   M   M   L   G   S   C   S   H   T   N   N   D   N   P   A
      .              .              .              .
      930             950             970
      .              .              .              .
TGGTGGCGGCTGGACCTAAAGAAAAGATACAAAGTGGACAAAGTGGTGATAGTGAACAGAGGAGACTGCT
      .              .              .              .
      W   W   R   L   D   L   K   K   R   Y   K   V   D   K   V   V   I   V   N   R   G   D   C   C
      .              .              .              .
      990             1010            1030            1050

```

GCGCTGAGAGACTGTTGGGAGCCCAGATTTCATATTGGAAATTCAGCAAATAACAACAACCCCAATATGCGG
A E R L L G A Q I H I G N S A N N N N P I C G
1070 1090 1110
CGGCATAAACAGCGTCTCTGAGGCCACTTACTCTGTCTGTCTGTCATGGGATGGAGGGTCAGTATGTGAGT
G I N S V S E A T I T L S C H G M E G Q Y V S
1130 1150 1170 1190
GTGGTCATTCCCGGGAGGGCAGAAAATCTCCAGCTCTGTGAAGTGGAGGTTTATGGGCAAGAAGTTAAAT
V V I P G R A E N L Q L C E V E V Y G Q E V K C
1210 1230 1250
GTGTTGCAGATAATTTGGCGAGACTGGGAGATGCCACTCAGAGCTCAACTTACAGACCCGAGTACAACGC
V A D N L A R L G D A T Q S S T Y R P E Y N A
1270 1290 1310 1330
TGGTGCCGCTATTGATGGCAATAAGGGGACAAATATGATGTTGGGCTCATGTTCCACACCAATAATGAC
G A A I D G N K G T N M M L G S C S H T N N D
1350 1370 1390
AACCCGGCTTGGTGGCGGCTGGACCTAAAGAAAAGATACAAAGTGGACAAAGTGGTGATAGTGAACAGAG
N P A W W R L D L K K R Y K V D K V V I V N R G
1410 1430 1450 1470
GAGACTGCTGCGATGAGAGACTGTTGGGAGCCCAGATTCTTATTGGAAATTCAGCAAATAACAACAACCC
D C C D E R L L G A Q I L I G N S A N N N N P
1490 1510 1530
AATATGCGGCGGCATAGACAGCGTCTCTGAGGCCACTTACTCTGTCTGTCTGTCATGGAATGGAGGGTCAG
I C G G I D S V S E A T I T L S C H G M E G Q
1550 1570 1590 1610
TATGTGAGTGTGGTCATTCCCGGGAGGGCAGAAAATCTCCAGCTCTGTGAAGTGGAGGTTTATGGGCAAG
Y V S V V I P G R A E N L Q L C E V E V Y G Q E
1630 1650 1670
AAGTTAAAGCTATTACAGCAGTGAATGTGGCAAGATGGGAAGCGTCAGTCAGAGCTCCACTTATAGACCA
V K A I T A V N V A R W E A S V R A P L I D Q
1690 1710 1730 1750
GAGTACAGCGCTGAAACAGCTATTGATGGTGATAAAGAGACAAATATCTTCATGCACCCGTGTGCCACA
S T A L K Q L L M V I K R Q I S S C T R V P T
1770 1790 1810
CTAATCCTGATAATCCTGCTTGGTGGCAACTGGACTTAAAGACTGCCTATATGATTGAGTCAGTGGTCAT
L I L I L L G G N W T * R L P I * L S Q W S *
1830 1850 1870 1890
AGTGAACAGAGGAGACTGTTGCAGTGAGCGCTGCTGGGAGCCCAGATCCGTGTTGGAAACTCACCATT
* T E E T V A V S A C W E P R S V L E T H H F
1910 1930 1950
CACAAACACCTGTATGCGGCACCATCACCGACGTCTCTGAAACGACCATCACTCTGTCTGTGACAAGA
T T T L Y A A P S P T S L K R P S L C P V T R
1970 1990 2010 2030
TGGAGGGTCGCTATGTGAGTGTGGTGATTCCCGGGAGAGCGGAATATCTCCATATCTGTGAAGTGGAGGT
W R V A M * V W * F P G E R N I S I S V K W R F
2050 2070 2090

```

TTATGGGGTGAAAATCTAAATATGATGCCCCACATGTTGGTTCAAGAATAGACATTTATTCATTATTAG
  M G * K S K Y D A P H V G S R I D I Y S F I R
    2110                2130                2150                2170
      .
      .
      .
      .
      .
      .
GCCGGGCATGGGTTATTGATTCAGTGTAATAATTGTGTTTAAGTTCGGTGCAATATTCTAGTCGTTTCTG
  P G M G Y * F S V K L C L S S V Q Y S S R F L
    2190                2210                2230
      .
      .
      .
      .
      .
      .
TTCATGACACTCAATATAAATAATCAATAAACTGAGATGTGAGAAAAAAAAAAAAAAAAATCGATATCTAG
  F M T L N I N N Q * T E M * E K K K K N R Y L D
    2250
      .
      .
      .
      .
      .
      .
ATCTGCGGCCCGCATGCGGA
  L R P H A

```

D. Xla-neurula cDNA

```

      10              30              50              70
      .              .              .              .
GAGAAAGAAGTGCAACACTAACAAGACSRACKAAGGAGAAAGAAGTGCAACACTAACAAGACCAACTGAC
R K K C N T N K T X X G E R S A T L T R P T D
      90              110              130
      .              .              .              .
AGGATGAAGTGCATTGTTCTCCTGCTGGTTTGCTTCTCTATCGGATGGGTTCACTCCAACCCCAAAAA
R M K C I V L L L V C F S I G W V H S N P T K K
      150              170              190              210
      .              .              .              .
AAGTTAACATTGCAAAATTTGGAGAAGCCTCACAGAGCTCAGATTACAGACCTGAGTACAATGCTGCTGC
V N I A K F G E A S Q S S D Y R P E Y N A A A
      230              250              270
      .              .              .              .
TGCTATCGATGGTGATAGAGACTCAAATATGATGGCGGGTTCATGCTCCCTTACTGGTAACGACAAGCCA
A I D G D R D S N M M A G S C S L T G N D K P
      290              310              330              350
      .              .              .              .
TCTTGGTGGCAGTTGAACCTAAAGCACAGGTACAAAGTGGAGAAGGTGGTGATAGTGAACAGAGGAGACT
S W W Q L N L K H R Y K V E K V V I V N R G D C
      370              390              410
      .              .              .              .
GCTGCAGTGAGCGCCTTTTGGGAGCCCAGATCCGTGTTGGATTACAGCCAATCTGAAGAACCCACTATG
C S E R L L G A Q I R V G F T A N L K N P L C
      430              450              470              490
      .              .              .              .
TGGCACCGTAACCTGATGTCTCTGAAGAAACCATCACTCTGTCCTGTCACGGGATGGTGGGTGAGTACGTC
G T V T D V S E E T I T L S C H G M V G Q Y V
      510              530              550
      .              .              .              .
ACTGTGTCTATTCTGAACGTGAGGAATATCTCCAGCTCTGTGAAGTGGAGGTCTATGGGAACAAATACA
T V S I P E R E E Y L Q L C E V E V Y G N K Y S
      570              590              610              630
      .              .              .              .
GCCCTGTTGTACCAGTCCATGAGGAATCGGAGGAAGATGTACTGCAAGACATAGGCAATTTATATAAACA
P V V P V H E E S E E D V L Q D I G N L Y K H
      650              670              690
      .              .              .              .
TTAAAGAACACCTTGGTTTTAGGGATGTCGCGGACTGTTGCGCCGTGAAGTAGTTTGCGCGAACATCGAC
*
      710              730              750              770
      .              .              .              .
TGTTGCGGTCCGCCGAATGTTGCGGAACGTGCGCTTAGCTGGCGCTTATTTTTGCCATTTCTCACCAGA
      790              810              830
      .              .              .              .
CCAGCAGATACATGGCAGCCAATCAGGAAGCTCTCCCTCCTGGACCACCCCAACCCCTGGACCACTC
      850              870              890              910
      .              .              .              .
CCCYTCCATATATAAACTGAAGCCCTGCWKYGTTTTTTTCATTCTGCCTGTGTGTGCTTGAAGAGCTAGT
      930              950              970
      .              .              .              .
GTAGGGAGAGAGCTGTTGAGTGATTTGAGGGACAGTTGATAGTAACTTTGCTGGCTAGTAATCTACTTGA
      990              1010              1030              1050
      .              .              .              .

```


TACTGCTCTGTATTGTAGGGACAGAACTCTGCAGGGATTTGAGGGACAGTGAGTTTAGGTTAGTTAGCTT
1070 1090 1110
TGCTGGCTAGTAATCTACCTTCTACTGCAGTGCTCTGTATGTAGCTGCTGTGGGCACTGCTGCTGATCTC
1130 1150 1170 1190
TCATCTGCTGACTGCTGCCCTGTAACCCAATAGTCCTTGTAAAGACTGCTTTTATTTTCTTTTTTGT
1210 1230 1250
TTACTTTGCTACTATAAGAGCCCAGTGCTATTAGTCTAGCTGTGTTGGGGAGTGGGACTGGTGTGCTGCT
1270 1290
CCTCCTAGTAGNTCACCCTACCAGCACCAACCAGAGTCAAAATTG

E. *In silico* XLEST2 cDNA

```

      10              30              50              70
      .              .              .              .
AAGGATGAAGTGCATTGTGGTTCTGCTCGYWTTTGCAGCTGTTGGGTGGGCGCAGTTGTGCAACCCCCAG
R M K C I V V L L X F A A V G W A Q L C N P Q
      90              110              130
      .              .              .              .
ATAGGAGGACAAAATTTGGCAAGATCAGGAGGAGTCAAGCAAAGCTCCACCTACGCTCCTCAGTACACTG
I G G Q N L A R S G G V K Q S S T Y A P Q Y T V
      150              170              190              210
      .              .              .              .
TTGATAAAGCGATTGATGGCATAAAAAACACAAATACCTTTGTACAAGCATGCGCCATTACTGGATATGA
D K A I D G I K N T N T F V Q A C A I T G Y D
      230              250              270
      .              .              .              .
CAAAAACGCTTGGTGGCAGGTGGACCTGAAGAATTCCTACAAAGTTGGTTCTGTGGTCATAGTGAACAGA
K N A W W Q V D L K N S Y K V G S V V I V N R
      290              310              330              350
      .              .              .              .
GGAGACTGTTGTGCCGATCGTCTGAAAGGAGCCCAGATCCGTGTTGGAAATTCAGCAGATAATAACAACC
G D C C A D R L K G A Q I R V G N S A D N N N P
      370              390              410
      .              .              .              .
CAGTATGCGCCACCGTCACTGATGTCTCTCAACTCACCATCAATATGTGCTGTAAGGGGATGGTGGGTCA
V C A T V T D V S Q L T I N M C C K G M V G Q
      430              450              470              490
      .              .              .              .
GTATGTGAGTGTGGTCATTCTGGCCGCAATGAATATCTCCAGCTCTGTGAAGTTGAGGTTTATGGGGAG
Y V S V V I P G R N E Y L Q L C E V E V Y G E
      510              530              550
      .              .              .              .
GAAAATAAACCTGAAGAAAAACCTGAAGAAAAACAACCTTTGTTGGTAAAACCATGTTACATTTCAGTCAGT
E N K P E E K P E E K Q L C W *
      570              590              610              630
      .              .              .              .
GGCCTCAGCAGGTGAAGGCAAATCAAGCAAATCAAGCAGCAGATCTCCCATCATTGTTCAGTGGTTGCTAC
      650              670              690
      .              .              .              .
ACTAGAACTTTCAAAACGTTCTTGGGGAATACAATGAGCAGCACTTCAACAAGAGTCCAGGCAAAGACCA
      710              730              750              770
      .              .              .              .
ACATACCATCTCTAATGGATACTGTAGAGCAGGGGTAGCCACAAGTTATTGGATCCTGATCTAACAATCG
      790              810              830
      .              .              .              .
ATCTTAAGCTGAAGATCATGACAGCAAATGTTGTCCACAAACAGGTTAATTAGATCCACATTCTCACTGT
      850              870              890              910
      .              .              .              .
TGAGATTCTTTCAATACTTTTGGGTTTGTTCATTTAACCTTATCAATATGGGCTCATTGATCATTTGAA
      930              950              970
      .              .              .              .
CGGGTTTAAGCGCTGAAACCATATGGGACAAAACATTTCCTTCGCTGAGGAACCATCGGTGGAGTATTTGT
      990              1010              1030              1050
      .              .              .              .

```

```

GCCTGTGGTTTTAGTGGTGGGAAGCTGTGAGCCTCACCTTGGGCTTGAGCAACTCCAGCACCTTGTGGCCT
      1070              1090              1110
      .               .               .
CCAGTGTATTCAATTAGTTGTGCTCTGTTGTACAAC TGCAAAGCATTTCACATCCGGGCTCATAAATGGA
      1130              1150              1170              1190
      .               .               .               .
GGGTTTTGTGTTTGTGTTTGTGTTGCCTGAGGCCACTGACAGAATGTGCAACTGTTCTTCTCTTTCATCCT
      1210              1230              1250
      .               .               .
CTCTCTCTTAATTGGTTTGGTGCAATTTCTGTTTTTCGAGAATGTTTTTTATTTTATACAAGCAGCCAA
      1270              1290
      .               .
AATCAATAAATGAATATGATTGTGAWAAACAAAAAAAAAAAAAAAAAAAAA

```

F. DreI-FBPL cDNA

```

      10              30              50              70
      .              .              .              .
AACACATGTGAGTTGGGAGGAGAAAAACAAACACATGACATGGATGTAAAGTTAAGCTAATGTGTTTTTC

      90              110              130
      .              .              .
CTAAGAAACCCATTCACTCAAACATACCCTGATGAAAGCAGTCAGAAAGTGCTGGAAATGAACATGAGGC
      M N M R L
      150              170              190              210
      .              .              .
TTTTTTACAGGAGTCTACCTCTCTTACTGGGGTTCTTATCAGTTCAGACGAGTGCGGCTGGCACAGAAGT
      F Y R S L P L L L G F L S V Q T S A A G T E V
      230              250              270
      .
GAACATAGCAGGATGGGGAACAGCCACCCAGTCGACAATATATATGGACGGGCTTCTGTAAATGCTCTG
      N I A G W G T A T Q S T I Y M D G L P V N A L
      290              310              330              350
      .
AATGGGATAAGCCCCCCTGTACACACACTATAGTACAGACTCTGCCCTGGTGGAGACTAGACCTGCAGA
      N G I S P P C T H T I V Q T L P W W R L D L Q K
      370              390              410
      .
AGAGCTACAGCGTGAACAGAGTCTCCATCACTAACAGACTGGACTGCTGCAGYGAGAGGATCAACGGTGC
      S Y S V N R V S I T N R L D C C S E R I N G A
      430              450              470              490
      .
TGAGATTCGCATCGGAGACGTTTCCTTCAGATGTGTTTCAGCAACCCAGTATGTGCAGTAGTTTCTACCATT
      E I R I G D V P S D V F S N P V C A V V S T I
      510              530              550
      .
CCAGCAGGACAGACGTTTCAGCTACTCGTGTAATGGGATGCAGGGACGTTACGTGTTTGTGGATATTAATG
      P A G Q T F S Y S C N G M Q G R Y V F V D I N A
      570              590              610              630
      .
CACCTTCAAGCATTCTTACTCTCTGTGCTGTAGGAGTCTTTGTGGTTTTTCCAGATAATTTGGCAACAGG
      P S S I L T L C A V G V F V V F P D N L A T G
      650              670              690
      .
TAAAAACGTGATGCAGTCATCAACCTACAGCTCCTGGATTCTTGAACAGGCCATTGATTTCAATCCAGGA
      K N V M Q S S T Y S S W I P E Q A I D F N P G
      710              730              750              770
      .
TTATCAGATCCATCAATAGGGTGTTCCTCAACCAATAGTCAAAGTACCCATGGTGGAGGCTGGATCTGG
      L S D P S I G C S S T N S Q T D P W W R L D L G
      790              810              830
      .
GCCACATTTACCAAGTGAGTACAGTAGTGGTCACTAACAGACTAAAGTGTGTCCAGAACGAATAAACGG
      H I Y Q V S T V V V T N R L N C C P E R I N G
      850              870              890              910
      .
AGCCGAGATTACATCGGAAACTCTTTGGAGAACAACGGCAACAACAATCCCATATGTGCTGTGATTTCC
      A E I H I G N S L E N N G N N N P I C A V I S
      930              950              970
      .
AGTATTCCTGCTGGAGTTTCTGCCACCTTCGCCTGTAAACAAATATGGAGGGTCGATATGTGAGTCTGTTGA
      S I P A G V S A T F A C N N M E G R Y V S L L I
      990              1010              1030              1050

```

TTCGTGGAGACACAAAGTTTCTCACTCTGTGTGAAGTGGAGGTCTATGGACAAGGTCCATGTTTGAAGCA
R G D T K F L T L C E V E V Y G Q G P C L K Q
1070 1090 1110

GTCATTGATGAAGCTGAAGCTCAACTCCAGCTTTAGTCTTTCTGAAGTTCGTCTCATGACACAGCTGGAA
S L M K L K L N S S F S L S E V R L M T Q L E
1130 1150 1170 1190

TCTGCTCTGGCACAAGAGGCTTTTCTGATGTCACGCTGCAATGGACACAAGTCCCAAACAGGAAGTGA
S A L A Q R G F S D V T L Q W T Q L P K Q E V I
1210 1230 1250

TACGGAAGAAAGTAGAGCAAGCTCACTGTGCTCAAACAAAGAGATGAAGGAAACGCTGAACTCTGGATCT
R K K V E Q A H C A Q T K R *
1270 1290 1310 1330

GCTGTGATCTTCATTCACATGCAAAAAATACATTAAATACAAACACGTGACCAACTGGAGCGACACGATT
1350 1370

CCGGAGAAACTGGGAATGAAATAAAAAATCTGACAAAATGCCAAAAAAAAAAAAAAAAA

G. DreII-FBPL partial CDS

```

ATAATCTGGCGCTAAAGAGAAACGCCTCTCAGTCGTCCAGTTTTGGGTTTTGGTCTGCAG
1  -----+-----+-----+-----+-----+-----+ 60
   N L A L K R N A S Q S S S F G F W S A E -

AAAGAGCCGTCGATGGTTCTCGAACTGGGCCAAAGTTATGGTCAGTGTGTTTCGTCCACCG
61  -----+-----+-----+-----+-----+-----+ 120
   R A V D G S R T G P K L W S V C S S T A -

CTAACCAAAGTAACCCGTGGTGGAGGGTGGATCTGGGGGACGTTTACCGCGTTAGCAGAG
121 -----+-----+-----+-----+-----+-----+ 180
   N Q S N P W W R V D L G D V Y R V S R V -

TAATCATCACTAACATTAACAACAGCGTCGCAGACCGAATAAACGGAGCCCAGATTCACA
181 -----+-----+-----+-----+-----+-----+ 240
   I I T N I N N S V A D R I N G A Q I H I -

TCGGAAACTCTCTGGAGAACAACGGCACCAACAATCCCATGTGTGCTGTGATTTCCAGTA
241 -----+-----+-----+-----+-----+-----+ 300
   G N S L E N N G T N N P M C A V I S S I -

TTCCAGCTGGAGTTTCTGCAACCTTCCTCTGTTGTTTATGGAGGGTCGATATGTGAGTC
301 -----+-----+-----+-----+-----+-----+ 360
   P A G V S A T F L C C F M E G R Y V S L -

TGTTTCATTCCCGGAGACTCAAAGATGCTYACTCTGTGTGAAGTGGAGGTGTATGTAGAAG
361 -----+-----+-----+-----+-----+-----+ 420
   F I P G D S K M L T L C E V E V Y V E G -

GTCCGTGTTGGAAGCAGTCGTTGGTAAACTGAAGCTCAACTCTAGCTTTAGTCTTTCTG
421 -----+-----+-----+-----+-----+-----+ 480
   P C W K Q S L V K L K L N S S F S L S E -

AAGTGCAGTTTCTGACCCAGCTGGAATCTGCTCTGGCCCGAAGAGGACTTTCTGACGTGA
481 -----+-----+-----+-----+-----+-----+ 540
   V Q F L T Q L E S A L A R R G L S D V T -

CGCTGCACTGGACGCAACTGCCCAAACAGATGCCAGAGCAAGTTGCGCCAGCTCAAAGAA
541 -----+-----+-----+-----+-----+-----+ 600
   L H W T Q L P K Q M P E Q V A P A Q R K -

AGAAACGCTGAAAACGAAACTCCCTGAGGATCAGCTTGGATTTTCTTCATCATTTAGTG
601 -----+-----+-----+-----+-----+-----+ 660
   K R *

CATTTATATTGCTTCAATATTTCTGTTTGTGTTGAATGTAGATCAAATATTTAATGTGATT
661 -----+-----+-----+-----+-----+-----+ 720

TTAATATTTACTGAATTTTATTGCTTTATCAGTTGTGCAGTTATCATTAATATTGTAAAT
721 -----+-----+-----+-----+-----+-----+ 780

AAATACACCAAGTATTTAAGCAAAAAAAAAAAAAAAAAAAAA
781 -----+-----+-----+-----+-----+

```

H. DreIII-FBPL CDS

```

ATGTTGTGGCTCGCTCTGCTCTTGGGGTTGTGTGTCACTGATGTGGCGCCGGCTAATCTG
1  -----+-----+-----+-----+-----+-----+ 60
  M  L  W  L  A  L  L  L  G  L  C  V  T  D  V  A  P  A  N  L  -

GCTCTGGGTGCCGCAGCGGTGCAGTCGTCCACTGGAGACCCCAATGGAAACGCTGAGCAC
61  -----+-----+-----+-----+-----+-----+ 120
  A  L  G  A  A  A  V  Q  S  S  T  G  D  P  N  G  N  A  E  H  -

GCAGTGGACGGAAACACAGAAGCAGACTACCGTAAGGGCTCCTGCACACACACCAGCCGT
121 -----+-----+-----+-----+-----+-----+ 180
  A  V  D  G  N  T  E  A  D  Y  R  K  G  S  C  T  H  T  S  R  -

GAGTTTAACCCCTGGTGGAGGGTGGATCTGGGTGGAGTGTCCAGCGTCAACAAAGTCACC
181 -----+-----+-----+-----+-----+-----+ 240
  E  F  N  P  W  W  R  V  D  L  G  G  V  S  S  V  N  K  V  T  -

ATCACCAACAGAGGAGACTGCTGTGAGGAGCGCATACGTGGAGCGCAGATCCGCATCGGG
241 -----+-----+-----+-----+-----+-----+ 300
  I  T  N  R  G  D  C  C  E  E  R  I  R  G  A  Q  I  R  I  G  -

GACAGCCTGGAGAACAACGGAACAACAACCAGCTAGCTGCTACTCTTCTGGATGCTATT
301 -----+-----+-----+-----+-----+-----+ 360
  D  S  L  E  N  N  G  N  N  N  Q  L  A  A  T  L  L  D  A  I  -

AAAGGCTCTCAGACGTTTCGAGTTTCAGCCGATTCAGGGCCGATACCTCAATGTTTTCTG
361 -----+-----+-----+-----+-----+-----+ 420
  K  G  S  Q  T  F  E  F  Q  P  I  Q  G  R  Y  L  N  V  F  L  -

CCCGGAATGATGAACTCTGAGTCTGTGTGAAGTGGAGGTGTTTTCAGCAGGTCCATCA
421 -----+-----+-----+-----+-----+-----+ 480
  P  G  N  D  E  T  L  S  L  C  E  V  E  V  F  S  A  G  P  S  -

AAGAATATCGCCGCTGGAGCCGCCCGCTTCAGTCCTCCACTTGGCCTCATGATGGTGAC
481 -----+-----+-----+-----+-----+-----+ 540
  K  N  I  A  A  G  A  A  A  V  Q  S  S  T  W  P  H  D  G  D  -

GCAGGAATGCTGTGGATGGAAGCAGCGAGTCTGAGTATCAGGAGGGCTCCTGTTACAC
541 -----+-----+-----+-----+-----+-----+ 600
  A  G  N  A  V  D  G  S  S  E  S  E  Y  Q  E  G  S  C  S  H  -

ACTCTGGGGGAAACCAACCCGTGGTGGAGGGTGGATCTAGGGAGAGTGTTCAGCATCCGC
601 -----+-----+-----+-----+-----+-----+ 660
  T  L  G  E  T  N  P  W  W  R  V  D  L  G  R  V  F  S  I  R  -

CGTGTGAGCATCACCAACAGAGGAGACTGCTGTGAGGAGAGGTTAAACGGAGCTGAGATC
661 -----+-----+-----+-----+-----+-----+ 720
  R  V  S  I  T  N  R  G  D  C  C  E  E  R  L  N  G  A  E  I  -

CGCATCGGGAACAGCCTGGAGAACAACGGAACAGCAACCACCTAGTTGCGACTGTGAG
721 -----+-----+-----+-----+-----+-----+ 780
  R  I  G  N  S  L  E  N  N  G  N  S  N  H  L  V  A  T  V  E  -

CACATTCAGCTGGAAACACCGAGACGTTTGAGTTCCAGCCGGTTCAGGGCAGATTCCTC
781 -----+-----+-----+-----+-----+-----+ 840
  H  I  P  A  G  N  T  E  T  F  E  F  Q  P  V  Q  G  R  F  L  -

```

AACATTGTACTGCCTGGTGTAACGTTTACCTCACTCTGTGTGAAGTGCAGGTGTTTACA
 841 -----+-----+-----+-----+-----+ 900
 N I V L P G V N V Y L T L C E V Q V F T -

GACTGA
 901 ----- 906
 D * -

I. DreIV-FBPL *in silico* CDS

```

AAATCTGGCTTTAAGTGGCAGGGCCACACAATCAGACCTGTTGAAAAATCCATGGACAGG
1  -----+-----+-----+-----+-----+-----+ 60
   N L A L S G R A T Q S D L L K N P W T G -

AGAGGCCCTTGCCAGTAATGCTATTGATGGAAATCGTGACCCAGATTTTTACCATGGGTC
61  -----+-----+-----+-----+-----+-----+ 120
   E A L A S N A I D G N R D P D F Y H G S -

TTGTACTGCCACTGAAGTACAAGATGATCCGTGGTGGAGGTTAGATTTACTAGACACATA
121 -----+-----+-----+-----+-----+-----+ 180
   C T A T E V Q D D P W W R L D L L D T Y -

TGTGGTGAAATCCATAACAATAACAAACCGAAAAGACTGCTGTCCTGAAAGACTTGATGG
181 -----+-----+-----+-----+-----+-----+ 240
   V V K S I T I T N R K D C C P E R L D G -

AGCCGAGGTTACATTTGGCAACTCTCTGCTGAACAATGGCAACAGCAATCCACTGGCTGC
241 -----+-----+-----+-----+-----+-----+ 300
   A E V H I G N S L L N N G N S N P L A A -

AAAAATTTCTCAATTCCAGCTGGAAGATCCCTCACTTTCAAATGGAAAAAAGGCATTTTC
301 -----+-----+-----+-----+-----+-----+ 360
   K I S S I P A G R S L T F K W K K G I S -

AGGTCGTTACATCAATGTAATCCTTCGTGGCTCCAATCAAATTCTTACCCTCTGCGAGCT
361 -----+-----+-----+-----+-----+-----+ 420
   G R Y I N V I L R G S N Q I L T L C E L -

TGAAGTTTACGGTTATCCGGCTCCCAATGGTGAAAATGTAGCTTTAAGAGGCAAAGCTAC
421 -----+-----+-----+-----+-----+-----+ 480
   E V Y G Y P A P N G E N V A L R G K A T -

ACAGAGTTTCTCTATGGAAATGGTTTTCGATCTAATGCAAATGATGGGAACAAAGATGG
481 -----+-----+-----+-----+-----+-----+ 540
   Q S F L Y G N G F A S N A N D G N K D G -

TGTTCACTCATGGATCCTGCACACACACTCACAAAACCTCAACCCCTGGTGGAGACT
541 -----+-----+-----+-----+-----+-----+ 600
   V H T H G S C T H T H K T L N P W W R L -

GGACCTGCTGAAAAGGCACAAAGTGTTTTCGGTGGTCATTACAAACACATTAGACAATCT
601 -----+-----+-----+-----+-----+-----+ 660
   D L L K R H K V F S V V I T N T L D N L -

TCCTGAAAGATTAAATGGTGCAGAAATACGAATTGGAGACAATCTGGACAACAATGGCAA
661 -----+-----+-----+-----+-----+-----+ 720
   P E R L N G A E I R I G D N L D N N G N -

TAATAATCCCAGATGTGCTACAATCGCTTCCATTCCTGCTGGTTTCTCCAGCTCTTTTGA
721 -----+-----+-----+-----+-----+-----+ 780
   N N P R C A T I A S I P A G F S S S F D -

CTGTGATGGGATGGAAGGACGTTACGTTAATGTTGTTATTCCAGGACGAGAAGAATATCT
781 -----+-----+-----+-----+-----+-----+ 840
   C D G M E G R Y V N V V I P G R E E Y L -

```

AACACTGTGTGAGGTTGAAGTTTACGGATCGCCACTGGACTGA
841 -----+-----+-----+-----+----- 883
T L C E V E V Y G S P L D * -

J. DreV-FBPL cDNA

```

      GGGATCGACACCGCAGGATGACGGTGGTGATGCTGGCTGTGCTGCTGCTGCTGCTGGTGTGTGTGTGTGA
1  -----+-----+-----+-----+-----+-----+-----+ 70
      M T V V M L A V L L L L L V C V C D

      CTCTACAGCCTTCAGACTCACAGGAAACATCGCTCTGAGAGCTGAAACCCATCAGTCCGCTGACCCGCTG
71  -----+-----+-----+-----+-----+-----+-----+ 140
      S T A F R L T G N I A L R A E T H Q S A D P L

      AACGGAGACTCCGCATGGAGAGCAGTGGCTGGAGATGGAGATCGGTCCTGCAGCTCCATCAGCTCGAAGA
141 -----+-----+-----+-----+-----+-----+-----+ 210
      N G D S A W R A V A G D G D R S C S S I S S K R

      GAAGCCCTGGTGGAGAGTTTCTCTAGCACAGACCTACAGAATCGCCAAAATCTCCATCAGCACCGGCAC
211 -----+-----+-----+-----+-----+-----+-----+ 280
      S P W W R V S L A Q T Y R I A K I S I S T G T

      TGAGGGCATCAGCGGGGCGAGATCCGCATTGGCAGCAGTCTGGAGGAGGACGAAACCACAACCAGCTG
281 -----+-----+-----+-----+-----+-----+-----+ 350
      E G I S G A E I R I G S S L E E D G N H N Q L

      GTGAGAGTGTCTCTGTGCGGCCGGGAAAGCTCAGGTGTTTAAAGTTCAGGCCTGTGGAGGGACGCTTCA
351 -----+-----+-----+-----+-----+-----+-----+ 420
      V R V F S V R P G K A Q V F K F R P V E G R F I

      TCACTGTGATTTTACCAGGAGTGGACCGTGTGCTGAATCTGTGTGAGGTGGAGGTGTTTCGCGCTCGCTGA
421 -----+-----+-----+-----+-----+-----+-----+ 490
      T V I L P G V D R V L N L C E V E V F A L A E

      AGACTCTAACTCAGACTCAGAGCTGGTGAATGTGGCAGTATCGGGTCGAGCCACTCAGTCCAGCATGCGT
491 -----+-----+-----+-----+-----+-----+-----+ 560
      D S N S D S E L V N V A V S G R A T Q S S M R

      CTGGGTTCTGCTGCGTGTCTCAGTCTGCCACAGAACGCCATCGACGGGAACCGGCAGTACGACCCGTCCC
561 -----+-----+-----+-----+-----+-----+-----+ 630
      L G S A A C L S L P Q N A I D G N R Q Y D P S R

      GCGGCTCATGTGCGCAGACAGACACCGAGAGCGCCCCCTGGTGGAGGCTGGACCTGCTGCGGACACACAC
631 -----+-----+-----+-----+-----+-----+-----+ 700
      G S C A Q T D T E S A P W W R L D L L R T H T

      CATCACTGCAGTGGCTCTGACCCGCGGAGACCAGGACGTCAACGGCGGAGGGTGACCATCGGGGACTCG
701 -----+-----+-----+-----+-----+-----+-----+ 770
      I T A V A L T R G D Q D V N G A R V T I G D S

      CTGCAGGATGAGGGACGAGCAAACCCGCTGTGTGTGTCCGTGTCTTCATCCCTGCTGGAGGCACAGGCT
771 -----+-----+-----+-----+-----+-----+-----+ 840
      L Q D E G R A N P L C V S V S F I P A G G T G C

      GCTTCAGGTGTGTTCTCTCTCCGGGGCCGATACGTGACTGTGGCTCTGGCTGGAGTAAACAGGACCCT
841 -----+-----+-----+-----+-----+-----+-----+ 910
      F R C V P A L R G R Y V T V A L A G V N R T L

      GAGCCTGTGTGAGGTGGAGGTGTTCTGGGGTGCTGATCAATAACATCCAGACATCAGCAGCGCTCCATAT
911 -----+-----+-----+-----+-----+-----+-----+ 980
      S L C E V E V F G V P D Q *

      TTACTGGCTCCTGTTTTTCATTTCTGCTTGTGTTGCATTATGGGATCTTGATCTTTCTTCCAACAACCTTTT
981 -----+-----+-----+-----+-----+-----+-----+ 1050

      AACCTTAAACAGTGTGTGAAAGTGTGTCCTGCACATGTGATAGTGTGCAGTGTCTATTGTGTGTGTGTG
1051 -----+-----+-----+-----+-----+-----+-----+ 1120

      TTGTGTCATAGATATATACACTAGATATCGCATAGAGACCCTGAGTATTCGTCTATAGCGCCGCCACATT
1121 -----+-----+-----+-----+-----+-----+-----+ 1190

```

```

GGTACAGTGCTCCCAGGACAAGTGTCAATCGCACTAGTCAAGACAGTGTTATTACGTGAAGATGCG
1191 -----+-----+-----+-----+-----+-----+-----+ 1260

GGACTTTAGCGCTGTCTATGAGTGTAGTAATGAGCAAAACAAAGAAAACAAAGCACAAAGGCAGAACACTT
1261 -----+-----+-----+-----+-----+-----+-----+ 1330

CATAGGTAATATTCAAGTTTTTCTGTTTTTGTATGTTCTGAACTCTTGTGCTAATCAGGTAACATTATTG
1331 -----+-----+-----+-----+-----+-----+-----+ 1400

ATGATAACAGTCACTTACTGCATTCACCATACGGCAAAGCAGCTCCAACTCGCACTAAACACTCGACTAA
1401 -----+-----+-----+-----+-----+-----+-----+ 1470

TGCTGGTTTTTGTGAATAAAATCAGCAAAACATGCAGGAGAAATATGACAACGAGATGCTGCGCTGCCAG
1471 -----+-----+-----+-----+-----+-----+-----+ 1540

AAACTTGTATTATTGTGCGCTAACGTTAGTGAAAGAGTCGTTTCGGGGGATTCATTACAAAGCGAGTCGCT
1541 -----+-----+-----+-----+-----+-----+-----+ 1610

CCTCCGTGAGCATGAGCAGTGACAGCAGGAGAGGAGCTGTGTTTCAGGACATGATTAGATCACATTTAAC
1611 -----+-----+-----+-----+-----+-----+-----+ 1680

AGGGAGGGTGAATAGTACATTTCTGTACACACAAACACAAGCTTCTGTGTCAGGAATGCCCGTGGGTGTC
1681 -----+-----+-----+-----+-----+-----+-----+ 1750

TGATCCATTAATGTAGAAAAGTGATGTAAACTATAATTTTCTTAATTAGAAAAAAATTATATACTAA
1751 -----+-----+-----+-----+-----+-----+-----+ 1820

CCTTCAGGAAACTCCCGATCATAGATATATGTGTATATGTGTATATCTCTGGCTTTGCATGGCCACAGT
1821 -----+-----+-----+-----+-----+-----+-----+ 1890

CCTCCACTGTACCTTGGTCCTGCATTCATTTCAAAGGAGCGCTACCCTGTAGCAGGATGGCGGCGCTATT
1891 -----+-----+-----+-----+-----+-----+-----+ 1960

GACGCATTCGGTCCAATAGACAATAACAGGCCAGGCGACATCTAGTGTATATATCTATGGTGTGTGTAT
1961 -----+-----+-----+-----+-----+-----+-----+ 2030

TACGCTACAGAAGCCTCGTGGTAAACTGCAGCACATTCACTGATACTCTACACACTCATTCCTCTGCAGA
2031 -----+-----+-----+-----+-----+-----+-----+ 2100

TGATCTCTTATATTATTTTAAACTGCTAGAAGTCTCTCTGGTAAACTGTAGAGTTGACTTTACAACATCA
2101 -----+-----+-----+-----+-----+-----+-----+ 2170

GCAGTGGACAGAGCAGATCAACATCCCATAATGCACTTCACCATCATAAATAAACACTTGACTAAATATG
2171 -----+-----+-----+-----+-----+-----+-----+ 2240

GACTAAATATGTGCTGTTTAAACAGTGAACAAACCAAATAACAGCGACGCTATTGACATTATTATAATAA
2241 -----+-----+-----+-----+-----+-----+-----+ 2310

ACAGCAAAAGCAGTCGCTCTGCAGACCTGAAGAGGTCCATGACCACATGACCAGCAGAACAAACATCATGA
2311 -----+-----+-----+-----+-----+-----+-----+ 2380

GAGGGAAGATCAGTTCACATCATCCCTGACCCAGACCTCCAATCATCAGTGGAAGAGAACGAGAGCTGAG
2381 -----+-----+-----+-----+-----+-----+-----+ 2450

GCCTGGAGACAGAGATGTTCAATCTGTTCCACAACAGCTGTTTAACTGAGAGATCCACTTTACTTCTCAA
2451 -----+-----+-----+-----+-----+-----+-----+ 2520

TATATTACTTTAATAGCTGATTTATTGTGTCTGTCTTAACCCTGTGTGTGTGTGTGTGTGTGTGTGTG
2521 -----+-----+-----+-----+-----+-----+-----+ 2590

TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG
2591 -----+-----+-----+-----+-----+-----+-----+ 2636

```

K. Omy-FBPL4 cDNA

```

      10              30              50              70
      .              .              .              .
TCTCCACATGGTGTCTCTGGAGCCTGTTACAGTTTACGGCTTCCCCCAGATTACCGTGTTAACTCCTCGC

      90              110             130
      .              .              .
TCCATGTGTCTCTCCTCAGGCAGGTGGTAGCTGGTCCACTCCAGGAGTCTGAGGTGTGGGAAGTTCCTCC

      150             170             190             210
      .              .              .              .
ACCCCCTTAGGACATCGAGGGGGCCACCGGCGTATGCAGTCCGATCCATACTGGACTTGACGTGTCAGGC

      230             250             270
      .              .              .
GAGGGGCTTTTCACTACCTCGTGGAGTGGGAGCGTGTAAGGTCCGGAGGAGAGATGCTGGGTGAGACAGCA

      290             310             330             350
      .              .              .              .
CAACCCCTGTCCCAGCCTCGGACGCATCCACCTCAACTATGAACGCTAAAGAGGGATCAGGATGAGCCAG

      370             390             410
      .              .              .
CACGGGAGCCGAGCAAGATGAAAGACGTAATTATACTCTCCCTGCTGGGGATTCTGGGAATGGGAAGGTG
      M K D V I I L S L L G I L G M G R C

      430             450             470             490
      .              .              .              .
CTCTAAAGATGAATTAGTGAATGTAGCTCTGAGGGGCAAAGCAGCCCAATCATCCACATCCTATGGAGGC
      S K D E L V N V A L R G K A A Q S S T S Y G G

      510             530             550
      .              .              .
ACTGCAAAAAGAGCCATTGATGGAATATGGAACCCGACTTACGAATATCTCTCTGCAGCCACACTTCGG
      T A K R A I D G I W N P T Y E Y L S C S H T S G

      570             590             610             630
      .              .              .              .
GAGAGACCAGCCCTGGTGGAGGGTGGACCTGCTGGAGACCTACCAAGTCACCTCCGTCACCATCACCAA
      E T S P W W R V D L L E T Y Q V T S V T I T N

      650             670             690
      .              .              .
CAGAGATACTTTAGCAGAGAGGATCAACGGGGCTGAGATCCGCGATTGGAAACTCCCTGGAGAACGATGGC
      R D T L A E R I N G A E I R I G N S L E N D G

      710             730             750             770
      .              .              .              .
AACAGTAACCCAGGTGTGCGTTGATATCCTCCATCCCCGCAGGGGGCTCCACCACATTTCAATGCCACG
      N S N P R C A L I S S I P A G G S T T F Q C H G

      790             810             830
      .              .              .
GGATGCGGGGGCAATACGTCAATGTGTTTCCTCAGAGGATACATGCAATATCTGACCCTGTGTGAGGTGGA
      M R G Q Y V N V F L R G Y M Q Y L T L C E V E

      850             870             890             910
      .              .              .              .
GGTGAATGCTCATCCTGCTCCTATTGAAATGGGACCAACTCAAGCTTCAGAACTCAACTTAATTGTTTCTT
      V N A H P A P I E M G P T Q A S E L N L I V P

      930             950             970
      .              .              .
GTTACAGACAATGTTGCCTTAAAGAAAACGACTCGCCAATCCTCCCAGTACTCCCACATGGGGGGCTCTA
      V T D N V A L K K T T R Q S S Q Y S H M G G S N

      990             1010            1030            1050
      .              .              .              .

```

ACAATGCCGTGGATGGAAGACGCCTCTCCATGTACAAAGACAAATCATGCAGCCGCACCAAGTCCCAAGT
 N A V D G R R L S M Y K D K S C S R T K S Q V
 1070 1090 1110
 CAACCCATGGTGGAGGGTGGACTTGCAGCGAGCGTACAACGTACCTCCATAACAGTCACAAACATTGAA
 N P W W R V D L Q R A Y N V T S I T V T N I E
 1130 1150 1170 1190
 GACGTTGATCCAGAGATGATCGATGGTGTGAAATCCACATTGGAAACTCGCTGCAGAATAATGGCAACA
 D V D P E M I D G A E I H I G N S L Q N N G N N
 1210 1230 1250
 ACAATCCATTGTGCGCTGTGATCTCCTCATATCCCGCGTGGGAGGTCATGACCTTTCAGTGCAGTGGGAT
 N P L C A V I S S Y P A W E V M T F Q C S G I
 1270 1290 1310 1330
 CGAGGGTCGCTATGTGAACGTTTTCTTGCCGGGCTGCAATAAACACCTGTCACTGTGTGAGGTGGAGGTG
 E G R Y V N V F L P G C N K H L S L C E V E V
 1350 1370 1390
 AATGTGGGTAGCCGTCCTGGTGAGGAGCAGACATTATATGATATGGACGAACATCTTTCCTCTGACAGGA
 N V G S R P G E E Q T L Y D M D E H L S S D R R
 1410 1430 1450 1470
 GGGTGCAAGATATGAAGAGAGGGCAAGACATTCTATGCCCAGACCATAAGACTCGATATGAGAATGCAGC
 V Q D M K R G Q D I L C P D H K T R Y E N A A
 1490 1510 1530
 CACTGGCGGGATAGCCAATCAGTCGTCGAGTGGGACATGTTTGGAGATGCCAACATGCCATTGACCTG
 T G G I A N Q S S Q W D M F G D A N N A I D L
 1550 1570 1590 1610
 AGCTGGAGCAACCGGTATCTGGAGGGTCTGCGAGCCACACGAAAGCAGAGGTCGACCCCTGGTGGCGGG
 S W S N R Y L E G S C S H T K A E V D P W W R V
 1630 1650 1670
 TGGACCTGAGCAAGACCCACAACGTCACCTACGTCACCATTACCAACAGAGGGGACTGCTGCTCAGACAG
 D L S K T H N V T Y V T I T N R G D C C S D R
 1690 1710 1730 1750
 GATCAGTGGAGCAGAGATCCACGTTGGGGACTCACTGTTCAACAACGGCAATAGCAACCCACTGTGTGCC
 I S G A E I H V G D S L F N N G N S N P L C A
 1770 1790 1810
 AGGATTCCATACATCCCTGCTGGGCAAAGCAGGACATTTCGCTGCGGTGGGATATCGGGGCGCTATGTCA
 R I P Y I P A G Q S R T F P C G G I S G R Y V N
 1830 1850 1870 1890
 ACATTCTCCTCCCAGGAAAAGAGAAGTACTTGACTTTGTGTGAGGTGGAAGTGCAAGCAAGCACCTTTCA
 I L L P G K E K Y L T L C E V E V Q A S T F Q
 1910 1930 1950
 GGCAGGCCTGCCTTCAACTGCCCACAAGACAGCTGTTACAGCTCCAAACCGAAACCTGGCCTGGTTGTGG
 A G L P S T A H K T A V T A P N R N L A W L W
 1970 1990 2010 2030
 GATCTTCTGATACCCCTCGGAAGAACTGCTATTTCGGAAAGAAGAGACCTATAATCAAAACAGTACAGAAG
 D L L I P L G R T A I R K E E T Y N Q N S T E E
 2050 2070 2090

AAGAGACATTTGATCATGACAAAATATCATTTGAGACACAACGTCCAGTTGATAAAGTGCCTGAGAATTT
E T F D H D K I S F E T Q R P V D K V P E N L
2110 2130 2150 2170
AGTCTCCGGTGGGAATAGAGGTCCATTCTCCAGTATGACAGCCACGGTGCTGCCAGCAATGCCACCGAC
V S G G I E V H S S Q Y D S H G A A S N A T D
2190 2210 2230
AGGAAGCGTAACCCACTGTACCATGCTGGCTCCTGCAGCCACACGGAGGCAGAGACCAACCCCTGGTGGA
R K R N P L Y H A G S C S H T E A E T N P W W R
2250 2270 2290 2310
GAGTGGACCTTTTGGACACATACCAAGTCACCTTTGTCAACCATCACCAACAGAGGGGACTGCTGTCTCCA
V D L L D T Y Q V T F V T I T N R G D C C L H
2330 2350 2370
CAAGATCAACGGAGCTGAGATCCGCATTGGAAACTCCCTGGAGAACAACGGGCACGACCAATCCACTGTGT
K I N G A E I R I G N S L E N N G T T N P L C
2390 2410 2430 2450
GCTGTGATCTCCGAGATGAGAGAAGGGCAGCCAATGGATATTCCATGTAACATGGAGGGACACTATGTCA
A V I S E M R E G Q P M D I P C N M E G H Y V T
2470 2490 2510
CTATCGTTCTCCAGGCAGAGAGAAGTACCTGGCACTGTGCGAGGTAGAGGTGTATGGGGGAAAGTGTATG
I V L P G R E K Y L A L C E V E V Y G G K *
2530 2550 2570 2590
ACMCTAACCAATCTGAAATGACTCCTTTTACAGGAAAGTTCCAAATAAGGCCAACATAGTTACCATCTAA
2610 2630 2650
CTACAGAATCACTTACTATTTTGAGAAGTGTGGAGAAAGTCACTCAGTATTTCTTGTTCATACCCAGTGGG
2670 2690 2710 2730
ATCTCAAATAGTTCTTTGTTGAAGGCGGACGGCGTATTTTAGCCAAATTATAAACAGTCACATTTGTATT
2750 2770 2790
CAGGTGTGTATTCTATTAATAAAGGCTTGTACTGTGAACCAAAAAAAAAAAAAAAAAAAAAA

L. CG9095 cDNA

The cleaved signal peptide is indicated by *lowercase* amino acids and are negatively enumerated. The protein motifs are indicated as follows: CCP are *underlined*; FBPL is *bold*; CTL is *boxed* and the transmembrane helix is *dotted underlined*. The in-frame stop codon is marked with an *asterisks*. The polyadenylation motif is *double underlined*.

```

1  GGACAGGAGCCGTGAGCGAGACAGCAACAATGCACCGCACGCAGCCGTCGCTGCCGCTGC
                                m h r t q p s l p l p   -19

61  CTCTGCCGCTGCTGGCGTTGGCGTTGGCGTTGGCTTCGCGCGCTGGCTTTTGCGCAGGCGC
    l p l l a l a l a l a s a l a f a q a Q   1

121 AGAATATAGATGCCGTTGCAGTTTCCCGGGATCGCCGGCGCACAGCAGCGTCGTCTTCT
    N I D A G C S F P G S P A H S S V V F S   21

181 CGAATGCGAATCTCACCCAGGGCACGGTGGCTCCTACAGCTGCGAGCGGGGATTTCGAGC
    N A N L T Q G T V A S Y S C E R G F E L   41

241 TTCTGGGACCGGCGCGCGTGTCTGCGACAAGGGGCAATGGGTGCCCGAGGGCATTCCGT
    L G P A R R V C D K G Q W V P E G I P F   61

301 TCTGCGTTTTGAACGTTGCCGCTGGCAAGGCGCCCATGCAGATTTCCACTGATGGCGCTG
    C V L N V A A G K A P M Q I S T D G A G   81

361 GTGCTCCACAAAAGGCCATCGATGGCTCCACATCCGCCTTCTTCACACCGGAGACCTGCT
    A P Q K A I D G S T S A F F T P E T C S   101

421 CGCTGACGAAGGCGGAGCGATCGCCCTGGTGGTATGTGAATCTCCTGGAACCCTACATGG
    L T K A E R S P W W Y V N L L E P Y M V   121

481 TGCAACTGGTGCGTCTGGACTTTGGCAAATCCTGTTGCGGCAATAAGCCCGCCACAATTG
    Q L V R L D F G K S C C G N K P A T I V   141

541 TAGTGCGAGTGGGCAACAACCGACCGGACTTGGGCACAAATCCGATCTGCAACCGCTTCA
    V R V G N N R P D L G T N P I C N R F T   161

601 CGGGCCTCCTGGAGGCCGGACAGCCGCTCTTCCTGCCCTGCAATCCCCGATGCCGGGAG
    G L L E A G Q P L F L P C N P P M P G A   181

661 CCTTCGTGAGTGTCCACCTGGAGAATAGCACACCCAATCCGCTGTCCATTGTGCGAGGCGT
    F V S V H L E N S T P N P L S I C E A F   201

721 TCGTCTACACGGACCAAGCGCTGCCCATCGAGCGGTGTCCCACCTTCCGCGATCAGCCGC
    V Y T D Q A L P I E R C P T F R D Q P P   221

781 CTGGAGCTCTGGCCTCGTACAATGGCAAGTGCTACATCTTCTACAACCGCCAGCCGCTGA
    G A L A S Y N G K C Y I F Y N R Q P L N   241

841 ACTTTTGGACGCACTGTCTCTGTGCGATCCCGTGGCGGTACGCTGATCAGTGAGAGCA
    F L D A L S F C R S R G G T L I S E S N   261

901 ATCCGGCGCTGCAGGGATTTCATCAGTTGGGAGCTGTGGCGGCGTCATCGCAGTGACGTCA
    P A L Q G F I S W E L W R R H R S D V S   281

961 GTTCGCAGTACTGGATGGGAGCGGTACGTGATGGCAGCGATCGCAGCAGCTGGAAATGGG
    S Q Y W M G A V R D G S D R S S W K W V   301

```


1021 TGAACGGTGACGAGCTGACCGTCTCCTTCTGGAGTCATCCCGGCGCGATGAGGATTGTG
 N G D E L T V S F W S H P G G D E D C A 321
 1081 CCGGATTTGATGGCTCCAAGGGCTGGCTCTGGAGCGATACCAACTGCAACACGCTGCTGA
 R F D G S K G W L W S D T N C N T L L N 341
 1141 ACTTCATCTGTGAGCACCACCGAAGACCTGTGGCCGACCGAGCAACCGCCCAATTCCA
 F I C Q H Q P K T C G R P E Q P P N S T 361
 1201 CGATGGTAGCCCTGAACGGATTTCGAGGTTGGCGCCAGATCAAGTACAGCTGCGATGCCA
 M V A L N G F E V G A Q I K Y S C D A N 381
 1261 ATCACCTGCTGGTGGGTCCCGCCACGAGGACCTGCCCTGGAGACTGGATTCTACAATGAGT
 H L L V G P A T R T C L E T G F Y N E F 401
 1321 TCCCGCCAGTGTGCAAGTACATCGAGTGTGGTCTGCCGCGCCAGCATTGCCCATGGTTCCT
 P P V C K Y I E C G L P A S I A H G S Y 421
 1381 ACGCCCTGCTCAACAACACGGTTGGCTACTTGAGCCTGGTGAAGTATTCGTGCGAGGAGG
 A L L N N T V G Y L S L V K Y S C E E G 441
 1441 GTTACGAGATGATAGGACGAGCTTTGCTCACCTGCGACTTTGATGAGCGCTGGAATGGAC
 Y E M I G R A L L T C D F D E R W N G P 461
 1501 CTCCACCACGTTGTGAGATTGTGGAGTGCACACTCTGCCCGGCAACTACTACAGCACCA
 P P R C E I V E C D T L P G N Y Y S T I 481
 1561 TTATCAACGCTCCCAATGGCACATACTACGGCTCCAAGGCGGAGATCAGTTGTCCACCCG
 I N A P N G T Y Y G S K A E I S C P P G 501
 1621 GATACCGCATGGAAGGACCTCGAGTGCTTACCTGCCTGGCCAGTGGTCAATGGAGCAGTG
 Y R M E G P R V L T C L A S G Q W S S A 521
 1681 CCCTGCCCGTTGCATCAAACCTGGAACCGTCCACTCAGCCCACTGCCGCGTCCACCATT
 L P R C I K L E P S T Q P T A A S T I P 541
 1741 CCGTGCCCTCGTCGGTGGCCACGCCACCACCGTTCCGCCCCAAGGTGGTCAGCTCGACCA
 V P S S V A T P P P F R P K V V S S T T 561
 1801 CCAGCCGCACCCCTACCGCCAGCAGTATCCACGGCGAGCAGCGGCATTGGCGGCAGCT
 S R T P Y R P A V S T A S S G I G G S S 581
 1861 CCACCAGCACAGTGGGCACGTATCCAGTCTCAGCCCCACGCAGGTGGAGATCAATGGCG
 T S T V G T Y P S L S P T Q V E I N G E 601
 1921 AATCTGAATCCGAAGAGGAAATCAATGTGCCTCCAGTGCCTGGCACCGTTTCGCGAGGAGT
 S E S E E E I N V P P V P G T V R E E F 621
 1981 TCCCACCACGACGCACAGTTCGTCCAGTGCTCATAACGAAGAAGCCGAACAGCACACCGG
 P P R R T V R P V L I P K K P N S T P A 641
 2041 CTGCCCTGCCGCCCACCACCCATCAGGTGCCACCGCAACCACCGTCCACCTACGCACCCA
 A L P P T T H Q V P P Q P P S T Y A P T 661
 2101 CACCACGCGCAGCTCGCGACCAAGTGGTGCTCCGAATAGCGCGGCGAAGTGGAGACAA
 P P R S S R P S G A P N S A G E V E T T 681
 2161 CCACGCGGAATACACAGCAGATCATCGCCAATTTCGCATCCGCAAGACAACGAGATCCCCG
 T R N T Q Q I I A N S H P Q D N E I P D 701

2221 ACAGTGTCAACATCCAGCAGAACCAGTCGCCCAATGTCAACGTGCCCTTCGCCGTCGATA
S V N I Q Q N Q S P N V N V P F A V D N 721
2281 ATCCCGACCGCAAGGAGACCAAGGAGGCCAAACTTAATCTGGGCGCCATCGTTGCTCTGG
P D R K E T K E A K L N L G A I V A L G 741
2341 GCGCTTTTGGTGGTTTCGTCTTCCTGGCCGCCGTCATCACAACGATCGTGATCCTTGTGC
A F G G F V F L A A V I T T I V I L V R 761
2401 GAAGAAACCGAACCACACAACACTATCGCCATCGCGCCTCGCCCGACTGCAACACTGTGG
R N R T T Q H Y R H R A S P D C N T V A 781
2461 CCAGCTTCGATAGCTCCACCTCCGGATCCCGCAATGGACTCAACAGGTACTACCGCCAAG
S F D S S T S G S R N G L N R Y Y R Q A 801
2521 CCTGGGAGAACCTGCACGAGTCCGCCTCGAAGAACAGCTCACACAACGCCCTCCGCCGCA
W E N L H E S A S K N S S H N A L R R K 821
2581 AGGAGACCCTCGATCCACCGAGCATGACCCGTTCCCGGGACAATCTGCGCGACAATATGC
E T L D P P S M T R S R D N L R D N M Q 841
2641 AGCGATCCCGCGAAAATCTCGACAGATGCGGCAGGGACAACACTACGGCATGCGGGATGACT
R S R E N L D R C G R D N Y G M R D D S 861
2701 CCGAGATGGTGGTGTCTCGGTGGTGTGCGATGTGTGCCTGAAGGGCGAGAAGAAGCGCC
E M V V S S V V S D V C L K G E K K R H 881
2761 ATCACCATCATCACCACAAGAGCAGCTCCCGCAACGCGGACTACCGCGATCGGGATCACT
H H H H H K S S S R N G D Y R D R D H S 901
2821 CCTCCGGCAGACGCGAGCACCACCGACATAGTGGTGGTGGTGGTGGCGCGGAGGCGGTG
S G R R E H H R H S G G G G G G G G G 921
2881 GTGGCGGCCACTATTGACCAACCATGGTCATCACCATCAATGTGGAGCGGCAGTAGCCAA
G G H Y * 926
2941 ACGAACGATAGTGGCCAGCAGCAGTCGGCGTAGAGATCGGATTCGGATTCGGACTTGGAT
3001 TTGGATTTTGATTGGACTCGGGTTTTTGGTTTTTGGATTTGGATTCAGATTCGAAAATCG
3061 CGATCTGAGAACTGCAATGCGAGCGCAACAACGAAACGTTTTTTGTTTAATTTTAGCATC
3121 AGTTTTTTTTTCGCGCATTAGTTATGTAAGCCACAGATGGAGAAAAAAGGGTTTCGGAAAA
3181 TGTAAGGAGAAAACTTTTCTGTTTATAGTGAAAAAATAAAAAAAAAAACTCGTATTAGGC
3241 CAGCCTATCCAACCCATTGCTCTGTGTCTAACACCAGGCTCTGTAAATATTCGATCCTA
3301 AGATTTACCTTAATGTATATTTAGTGACTTTCTTAGACCCGATCCCTTTTCGACTTTCCC
3361 CTCTTTCACCCAGTTTAGATCCCTCGCTTCTATGGTTATAGGTCGTCAGTTTTTCATTTAA
3421 AGTTTCTGTACAAACAATATCTTTCTCAATGTAAACACACAAAACTCGTATAATTAGAG
3481 TACACCTAAACTTAATTTATGGTTAATAAACGTTGATATTCAAAACCCAAAAAAAAAAAAA
3541 AAAAAA

M. Pairwise alignment of MsaFBP32 and MchFBP32 genes.

Exons are in uppercase. Introns are in lowercase. The Tc1/*mariner*-like element (TLE) is in lowercase and underlined.

```
>Msfbp      CTGAACTCCA GGGTTAAAAG ATCTGTTCTA ACCAGGAAGC AGGgtaagca
>Mcfbp      CTGAACTCCA GGGTTAAAAG ATCTGTTCTA ACCAGGAAGC AGGgtaagca
#1          .....
          -----

>Msfbp      ctttagttat ttcacatgat catctcgat tttacaaatt gttgtcattt
>Mcfbp      ctttagttat ttcacatgat catctcgat tttacaaatt gttgtcattt
#51         .....
          -----

>Msfbp      tgcacattta tttattttatt atttatatatt gtatagcacc aaatgataac
>Mcfbp      tgcacattta tttattttatt atttatatatt gtata::: :::::
#101        .....
          -----*****

>Msfbp      agatgagcta aaacgagtggt ttctcgggggt ccattgggtgc tatataaata
>Mcfbp      ::::: ::::: ::::: ::::: :::::
#151        .....
          *****

>Msfbp      atccacaaca aagaaaagaa atgcaaca:: :gtcattct gcactacaaa
>Mcfbp      ::::: ::gaaaagaa atgcaacaaa ccgtcattct gcactacaaa
#201        .....
          ***** **-----**

>Msfbp      cattaaaagc tggagtaaca ggaatgcggt gtgcacatga atataagatt
>Mcfbp      cattaaaagc tggagtaaca ggaatgcggt gtgcacatga atataagatt
#251        .....
          -----

>Msfbp      tatgataatt atgctgtaca gatatgtag tgaagaacta tgaatcgata
>Mcfbp      tatgataatt atgctgtaca gatatgtag tgaagaacta tgaatcgata
#301        .....
          -----

>Msfbp      acaaaaatac aacctaataga tacaatttta caaaaatact tatatcctat
>Mcfbp      acaaaaatac aacctaataga tacaatttta caaaaatact tatatcctat
#351        .....
          -----

>Msfbp      aattcgtgct ctctaaaatg tccaatagtt agtcccatgt tccccatact
>Mcfbp      aattcgtgct ctctaaaatg tccaatagtt agtcccatgt tccccatact
#401        .....
          -----

>Msfbp      cttgttttcc tcccctttac ccctaataatc tacattttcaa gcaacattat
>Mcfbp      c::gttttcc tcccctttac ccctaataatc tacattttcaa gcaacattat
```

```

#451      .....
          -**-----

>Msfbp    tttattatta tttttctaaa ttaaaaaaac tactttcaaaa tcattttgat
>Mcfbp    tttattatta tttttctaaa ttaaaaaaac tactttcaaaa tcattttgat
#501      .....
          -----

>Msfbp    agagcagctt ctataaagat agctactggc cgaggggaaa aaagttaaga
>Mcfbp    agagcagctt ctataaagat agctactggc cgagaggaaa aaagttaaga
#551      .....
          -----*-----

>Msfbp    ggacggcaga ggaacctaag aagagacatc agaaaagact acgtctcgtg
>Mcfbp    ggacaacaga ggaacctaag aagagacaga aga::gact acctctcgtg
#601      .....
          ----**-----**---***---*-----

>Msfbp    ctgggttttta gttgttggtg agagccgtgt gaaagaatag atcgtttcat
>Mcfbp    ctgggttttta gttgttggtg agagctgtgt gaaagaatag attgtttcat
#651      .....
          -----*-----*-----

>Msfbp    tcgtacttgc agcaaagtgt tgcagtaaac ttttaagagca gatcaaaaat
>Mcfbp    tcgtacttgc agcaaagtgt tgcagtaaac ttttaagagca gatcaaagat
#701      .....
          -----*-----

>Msfbp    gatcacatga tcaatcacat gttcaaagca gggatagcca atcagagact
>Mcfbp    gatcacatga tcaattacat gttcaaagca aggatagcca atcagagact
#751      .....
          -----*-----*-----

>Msfbp    taaccagcct cttttcgatt gttcattctt cttccttctc tttttgtttg
>Mcfbp    taaccagcct cttttcgatt gttcattctt cttccttctc tttttgtttg
#801      .....
          -----

>Msfbp    cagttagtct tcattcaaga taattgcgac aaagtcatgt catcttcttc
>Mcfbp    cagctagtct tcattcaaga taattgcgac aaagtcatgt catcttcttc
#851      .....
          ---*-----

>Msfbp    ttgaagactt tcatcagtca aaatgtactt ggagatgcca tttggagctt
>Mcfbp    ttgaagactt tcatcagtca aaatgtactt ggagatgcca tttggagctt
#901      .....
          -----

>Msfbp    taatgctttt aacaattctc tctcttgtct ttctcaacca gAATAATGAG
>Mcfbp    taatgctttt aacaattctc tctcttgtct ttctcaacca gAATAATGAG
#951      .....
          -----

>Msfbp    GCACAGTGTG GTATTCTCTG TGCTGCTCCT CTTAGGGGCG TGTTTCAGCTT

```

```

>Mcfbp      GCACAGTGTG GTATTTCTGT TGCTGCTCCT CTTAGGGGCG TGTTTCAGCTT
#1001
.....
-----

>Msfbp      ACAACTATAg taagtataat gtaaaaaata tcctgtattt acagttgcaa
>Mcfbp      ACAACTATAg taagtataat gcagaaaata tcctgtattt acagttgcaa
#1051
.....
-----_*_*-----

>Msfbp      tcagaaat:: :::::::::: :::::::::: :::::::::: ::::::::::
>Mcfbp      tcagaaatta aagctgcaag cagcggtggg cgggacctcg ccctccacgc
#1101
.....
-----** *****

>Msfbp      :::::::::: :::::::::: :::::::::: ::::::::::
>Mcfbp      aactcggggc gtgctgcggc cgcgccacgt gccgttgtaa cacttgcgta
#1151
.....
*****

>Msfbp      :::::::::: :::::::::: :::::::::: ::::::::::
>Mcfbp      ccacaacatg aaagtttcac gcaaatcaga tcaaagtcgc tatgcagtgc
#1201
.....
*****

>Msfbp      :::::::::: :::::::::: :::::::::: ::::::::::
>Mcfbp      ttacttctctg ttgccagtag gtggcgctat cactaaatct gaatattgtc
#1251
.....
*****

>Msfbp      :::::::::: :::::::::: :::::::::: ::::::::::
>Mcfbp      atgtagacgt gttcagggca ggactgttat caaacatgtg acgtttcatt
#1301
.....
*****

>Msfbp      :::::::::: :::::::::: :::::::::: ::::::::::
>Mcfbp      gagattggac catgtacacg gaagttgtga ggacgtactg aagataggac
#1351
.....
*****

>Msfbp      :::::::::: :::::::::: :::::::::: ::::::::::
>Mcfbp      tactatctgt tttatggcca atcgggtggaa ttcgacacat tgtcccgggc
#1401
.....
*****

>Msfbp      :::::::::: :::::::::: :::::::::: ::::::::::
>Mcfbp      acgcggttaa acgtacagtc acaaatagca caactgttga tcaccagtgt
#1451
.....
*****

>Msfbp      :::::::::: :::::::::: :::::::::: ::::::::::
>Mcfbp      gtctggattg cagtgatgga atttgaagtc aatctgacaa tatctgtagg
#1501
.....
*****

```

```

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      aggagttcat taaaacagaa gcccataata gacagaaatg gcatcacagt
#1551      .....
          *****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      ttgagattcg attcaacctg gccaaactttg acggcacgcc acagacacgc
#1601      .....
          *****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      ccttcaacgt acatacacia atagcacaaac tttttatcag caccattttt
#1651      .....
          *****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      tggctgtact gaccaatttt gaagcagatc taataaattg cctaggagga
#1701      .....
          *****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      gttcgttcaa atagaaggcg aaaaattaca gaaaatgacc aaaaaagaca
#1751      .....
          *****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      aaacggccga ctttgatgcc acgccacgga cacgcccttc aacgtacatg
#1801      .....
          *****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      cacaaatagc acaactttta atcagcacca ttttttgact gtactgacca
#1851      .....
          *****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      atattgaagc agatccgatg aattccctag gaggagttag tccaaatata
#1901      .....
          *****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      acgcgaaaat ttcaccgaaa gtgaatgaaa atcaaaatgg cggacttccc
#1951      .....
          *****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      gtgggggttg gggatatgggt ccaagaggct tttttgtacg tcttggtgag
#2001      .....
          *****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      ttacacgtgc ctaccgagtt tcgtacatat cggtgaaacg tggcgccggg
#2051      .....
          *****

```

```

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      gctgtttttt tttgttatcc tgcagggggc gctacaaacc cgacaaacct
#2101      .....
*****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      ccacttcctg ttttgaccaa aatgtacagg aagttttaat tgttttatgt
#2151      .....
*****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      tttttgccag ggcggtcccc aggtgctgcg ggaagaattt cgtgcaaatc
#2201      .....
*****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      ggacaaagcc tgtaggagaa acatgaaaaa gtagtttgag gacatttcgt
#2251      .....
*****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      gctctaacgg aaaaacattc taggcggaag tgggcgtggc ctatatgggg
#2301      .....
*****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      agattcagct ccattcaccc aacatgttat atatataat gatataaggt
#2351      .....
*****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      ttttgtatgt gcgacaaacc atgcagaagt tattagccta aacgcgtttt
#2401      .....
*****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      ttatgatagt agcgccacct agtggcaa atcgaaagcaa cacaaagart
#2451      .....
*****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      atcaaatttt tcaccaggcc tgaccacttt gccaaataaa attgagtttt
#2501      .....
*****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      cgtatatgtt caggggggca aaaatgcagt cgaagtcgct aaaaataata
#2551      .....
*****

>Msfbp      ::::::::::: ::::::::::: ::::::::::: ::::::::::: :::::::::::
>Mcfbp      ataggtaaga agaagaaaca tagcagtttc aatagggacc tccagcgatc
#2601      .....

```

```

*****
>Msfbp      ::::::::::: :::::::::::aa tcaaccccct cacctcagaa aacattatag
>Mcfbp      gctggctcag tccctaataa tcaaccccct cacctcagaa aacattatag
#2651      .....
*****
>Msfbp      tgacaataaa tgacaaaaaa cctcattaaa caa::::::::: ::::::::::tat
>Mcfbp      tgacaataaa tgacaaaaaa cctcattaaa caacaaaaaa tatttgatat
#2701      .....
-----
>Msfbp      attgataca tatgatttat ttgacaaca gaagttgact taactttgaa
>Mcfbp      atatga::::: :::::tttat ttgacaaca gaagttgact taactttgaa
#2751      .....
--*---*** *****
>Msfbp      gttcaaatga aagctgtaac tcttttaaca cgtgcagtat tattcaaccc
>Mcfbp      gttcaaatga aaactgtaac tctt::::::::: :::::::::::
#2801      .....
-----
>Msfbp      ccagcttcag tactttgtag cgcaccccaa cgtctaacaa acgttttttg
>Mcfbp      ::::::::::: ::::::::::: ::::::::::: :::::::::::
#2851      .....
*****
>Msfbp      gttctttcac ctctcgatca caatcttttc ccattgttca cgtgcaaaag
>Mcfbp      ::::::::::: ::::::::::: ::::::::::: :::::::::::
#2901      .....
*****
>Msfbp      tctctatctc agtgaatgtt gatggcttcc gtgctgcctt ctttaaattc
>Mcfbp      ::::::::::: ::::::::::: ::::::::::: ::aagttctt cttt:::::
#2951      .....
*****
>Msfbp      caccaaagtt aaatctagtg actgaacagg ccacttcagg atgttcagg
>Mcfbp      :::::::::::t aaatctggtg actgaacagg ccactccagg atgttcagg
#3001      .....
*****_-----*-----
>Msfbp      atctttttct caaccaagct ttggttgact tggagatgtg cttgggatca
>Mcfbp      atctttttct caaccaagct ttggttgact tggagatgtg cttgggatca
#3051      .....
-----
>Msfbp      ttgtcttgct ggaaagtcca ttgatcacta aggtttaata tgttaacaga
>Mcfbp      ttgtcttgct ggaaagtcca ttgatcacta aggtttaata tgttaacaga
#3101      .....
-----
>Msfbp      aggcacacg ttactcttta aaatggcctg gtatttctgg gaatccatga
>Mcfbp      aggcacacg ttactcttta caatggcctg gtatttctgg gaatccatga

```



```

#3151      .....
          ----- *----- *-----
>Msfbp      tgccaggtac acgatcaaaa ttcacaaaca gtctattatt atatctttta
>Mcfbp      tgccaggtac acgatcaaaa ttcacaaaca gtctattatt atatctttta
#3201      .....
          -----
>Msfbp      tategatatc aactgtgtcc ttcataagAAA ATGTGGCCTT GCGTGGAAAA
>Mcfbp      tategatatc aactgtgtcc ttcataagAAA ATGTGGCCTT GCGTGGAAAA
#3251      .....
          -----
>Msfbp      GCGACTCAGT CGGCACGTTA TTTGCACACA CATGGAGCCG CCTACAACGC
>Mcfbp      GTGACTCAGT CGGCACGTTA TTTGCACACA CATGGAGCCG CCTTCAACGC
#3301      .....
          ----- *-----
>Msfbp      CATTGATGGA AACCGTAACT CTGACTTCGA AGCTGGATCG TGCACCCACA
>Mcfbp      CATTGATGGA AACCGTAACT CTGACTTCGA AGCTGGATCA TGCACCCACA
#3351      .....
          ----- *-----
>Msfbp      CTATTGAACA GACCAACCCC TGGTGGAGAG TGGACCTACT GGAGCCCTAC
>Mcfbp      CTGTTGAACA GACCAACCCC TGGTGGAGAG TGGACCTACT GGAGCCCTAC
#3401      .....
          ----- *-----
>Msfbp      ATCGTCACCT CCATCACCAT CACCAACAGA GGAGACTGCT GTCCAGAAAG
>Mcfbp      ATCGTCACCT CCATCACCAT CACCAACAGA GGAGACTGCT GTCCAGAAAG
#3451      .....
          -----
>Msfbp      GCTCAACGGG GTGGAGATTC ACATCGGCAA CTCTATACAA GAAAATGGTG
>Mcfbp      GCTCGATGGA GCGGAGATTC ACATCGGCAA CTCTTTACAA GAAAATGGTG
#3501      .....
          ----- *-*-* *-----
>Msfbp      TTGCAAACCC AAGgtgagtg catattaaca gttataagtg aaaacagtga
>Mcfbp      TTGCAAACCC AAGgtgagtg catattaaca gttataagtg aaaacagtga
#3551      .....
          -----
>Msfbp      cagctagacc cacaattctg ttcaatgttc attacatttt acctttacct
>Mcfbp      tagctagacc cacaattctg ttcaatgtta attacatttt acctttacct
#3601      .....
          ----- *-----
>Msfbp      ttaacctcag tccagtcagt gcagtcacaa cagacctgtt tcctaaccat
>Mcfbp      ttaacctcag tccagtcagt gcagtcacaa cagacctgtt tcctaaccat
#3651      .....
          -----
>Msfbp      gacactacgt agatgaacat cattaatgtg ctgctttctc tcttcttctg

```

```

>Mcfbp      gacactatgt agatgaacat cattaatgtg ctgctttctc tcttcttgtg
#3701      .....
          -----*-----

>Msfbp      tctcttttag GGTGGGTGTA ATTTCTCATA TCCCTGCAGG GATCTCACAT
>Mcfbp      tctcttttag GGTGGGTGTA ATTTCTCATA TCCCTGCAGG GATCTCACAT
#3751      .....
          -----

>Msfbp      ACTATCAGTT TCACTGAACG TGTGGAGGGA CGTTACGTGA CTGTGCTTCT
>Mcfbp      ACTATCAGTT TCACTGAACG TGTGGAGGGA CGTTACGTGA CTGTGCTTCT
#3801      .....
          -----

>Msfbp      ACCTGGTACA AACAAAGTTC TTACTACTCTG TGAAGTGGAG GTTCATGGGT
>Mcfbp      ACCTGGTACA AACAAAGTTC TTACTACTCTG TGAAGTGGAG GTTCATGGGT
#3851      .....
          -----

>Msfbp      ACCGAGCCCC AACTGgtgag aatttgagtc ataccatatt gtatatttgc
>Mcfbp      ACCGAGCCCC AACTGgtgag aatttgagtc ataccatatt gcatatttgc
#3901      .....
          -----_*-----

>Msfbp      aatttaggta attctaatta tatcagtagc taaataatac agagatcaga
>Mcfbp      aatgaaggta attctactta tatcagtagc taaataatac agagatcaga
#3951      .....
          ---**-----_*-----

>Msfbp      gggtatccag cctgcaggtc ccacaagaac cacagcagat ttcagtgacg
>Mcfbp      gggtatccag cctgcaggtc ccacaagaac cacagcagat ttcagtgacg
#4001      .....
          -----

>Msfbp      gttgcccaac tcgccccaaaa tgattcttac tttgatcaca atttgatcca
>Mcfbp      gttgcccaac tcgccccaaaa tgattcttac tttgatcaca atttgatcca
#4051      .....
          -----

>Msfbp      ttccgtccca taacattttg aaaatctaga agagctgcat gctactaatc
>Mcfbp      ttccgtccca taacattttg aaaatctaga agagctgcat gctactaatc
#4101      .....
          -----

>Msfbp      ttggacgtct ttgatattta gctggactct ::::::::::: :::::::::::
>Mcfbp      ttggacgtct ttgatattta gctggactct ggtgtgcatt ctctacgggc
#4151      .....
          -----*****

>Msfbp      ::::::::::: cagaggcttt ttgtttgaac atagtattca gttctcttta
>Mcfbp      ccaaaaagat cagaggcttt ttgtttgaac atagtattca gttctcttta
#4201      .....
          *****-----

```

```

>Msfbp      acgtgcctta gctggtagga aaacaagggt gataaatact gtccacagtg
>Mcfbp      acgtgcctta gctggtagga aaacaagggt gatgaatact gtccacagtg
#4251      .....
            -----*-----

>Msfbp      ttaaattatt agagacaagc tcattgtatg agtgctgtag catggccccg
>Mcfbp      ttaaattatt agagacaagc tcattgtatg agtgctgtag catggccccg
#4301      .....
            -----*-----

>Msfbp      atgtaagtgt ttaattaatc taaacatcta tttgccacag tggcaaggcc
>Mcfbp      atgtaagtga ttaattaatc taaacatata tttgccacag tggcaaggcc
#4351      .....
            -----*-----*-----

>Msfbp      ttatactaat aaacctgtta ttagtataaaa caagcaacaa caaggtaaat
>Mcfbp      t::::::::: :::::gtta ttagtataaaa cacgcaa::a caaggaaaaat
#4401      .....
            _***** *****_-----*-----*-----

>Msfbp      tcaaagtgct cagtcgtgaa agcagaattg ctcttccata aagcgtgggg
>Mcfbp      tcaaagtgct cagtcgtgaa agcagaattg ctcttccata aagcgtgggg
#4451      .....
            -----

>Msfbp      cttgccagag acagattaca aacagtctag atgcagcaga gaacattaaa
>Mcfbp      cttgccagag acagattaca aacagtctag atgcagcaga gaacattaaa
#4501      .....
            -----

>Msfbp      caccaaaaac aacaaacaaa acctcaatat cagcgctgag accatatttt
>Mcfbp      caccaaaaac aacaaacaaa acctcaatat cggcgctgag accatatgtt
#4551      .....
            -----*-----*-----

>Msfbp      tgttgcagGA GAGAACCTGG CCCTCAAAGG AAAAGCCACA CGGTCGTCAT
>Mcfbp      tgttgcagGA GAGAACCTGG CCCTCAAAGG AAAAGCCACA CAGTCGTCAT
#4601      .....
            -----*-----

>Msfbp      TGT'TTGAATC TGGTATTGCA TATAATGCCA TTGATGGGAA TCAAGCCAAC
>Mcfbp      TGT'TTGAATC TGGTATTGCA TATAATGCCA TTGATGGGAA TCAAGCCAAC
#4651      .....
            -----

>Msfbp      AATTGGGAAA TGGCCTCCTG CACTCACACA AAAAACACAA TGAACCCCTG
>Mcfbp      AATTGGGAAA TGGCCTCCTG CACTCACACA AAAAACACAA TGAACCCCTG
#4701      .....
            -----

>Msfbp      GTGGCGAATG GATCTGAGCA AAACCCACAG AGTGTTTTCT GTTAAGGTAA
>Mcfbp      GTGGCGAATG GATCTGAGCA AAACCCACAG AGTGTTTTCT GTTAAGGTAA
#4751      .....
            -----

```

```

>Msfbp      CCAACCGAGA TTCATTTGAA AAACGAATCA ATGGAGCTGA GATCCGAATT
>Mcfbp      CCAACCGAGA TTCATTTGAA AAACGAATCA ATGGAGCTGA GATCCGAATT
#4801      .....
            -----

>Msfbp      GGAGATTCCC TCGACAACAA CGGCAACAAC AATCCCAGgt agtttactga
>Mcfbp      GGAGATTCCC TCGACAACAA CGGCAACAAC AATCCCAGgt agtttactga
#4851      .....
            -----

>Msfbp      ctgттаatta attgcaatac caatataatt aaatataaat gtggтаactt
>Mcfbp      ctgттаatta attgcaatac caatataatt aaatataaat gtggтаactt
#4901      .....
            -----

>Msfbp      tttttacatt atcgccataa tgaagtгatt tcttcttctt ctaccagGTG
>Mcfbp      tttttacatt atcgccataa tgaagtгatt tcttcttctt ctaccagGTG
#4951      .....
            -----

>Msfbp      TGCTGTGATC ACAAGCATCC CAGCAGGTGC TTCTACTGAA TTCCAGTGTA
>Mcfbp      TGCTGTGATC ACAAGCATCC CAGCAGGTGC TTCTACTGAA TTCCAGTGTA
#5001      .....
            -----

>Msfbp      ACGGGATGGA TGGCCGCTAT GTTAACATTG TTATCCCTGG AAGAGAAGAG
>Mcfbp      ACGGGATGGA TGGTCGCTAT GTTAACATTG TTATCCCTGG AAGAGAAGAG
#5051      .....
            -----*-----

>Msfbp      TACCTGACCC TGTGTGAGGT GGAGGTGTAT GGCTCTGTCC TGGATTAGGT
>Mcfbp      TACCTGACCC TGT
#5101      .....
            -----

>Msfbp      GTCAGTACTA ATACTGTTGA ATGTACACAA ACAAACAAA ATAGTAGATT
#5151      .....
            -----

>Msfbp      AAGCTTTTTT GATTGTTTCC ATTCAAAATA AGACAGAGAT GGTCTTATCC
#5201      .....
            -----

>Msfbp      AATAAAA
#5251      .....
            -----

```

BIBLIOGRAPHY

1. Kaiser, D., Building a multicellular organism. *Annu Rev Genet*, 2001. **35**(1): p. 103-123.
2. Du Pasquier, L. and M. Flajnik, *Origin and evolution of the vertebrate immune system*, in *Fundamental immunology*, W.E. Paul, Editor. 1999, Lippincott-Raven: Philadelphia. p. xxi, 1589 p.
3. Paul, W.E., ed. *Fundamental immunology*. 4th ed. 1999, Lippincott-Raven: Philadelphia. xxi, 1589 p.
4. Weintraub, A., Immunology of bacterial polysaccharide antigens. *Carbohydrate Research*, 2003. **338**(23): p. 2539-2547.
5. Bork, P., L. Holm, and C. Sander, The immunoglobulin fold. Structural classification, sequence patterns and common core. *J Mol Biol*, 1994. **242**(4): p. 309-20.
6. Medzhitov, R. and C.A. Janeway, Jr., Innate immunity: Impact on the adaptive immune response. *Curr Opin Immunol*, 1997. **9**(1): p. 4-9.
7. Carreno, B.M. and M. Collins, The B7 family of ligands and its receptors: New pathways for costimulation and inhibition of immune responses. *Annu Rev Immunol*, 2002. **20**(1): p. 29-53.
8. Carroll, M.C., The role of complement and complement receptors in induction and regulation of immunity. *Annu Rev Immunol*, 1998. **16**: p. 545-68.
9. Janeway, C.A., Jr., Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harb Symp Quant Biol*, 1989. **54**(Pt 1): p. 1-13.
10. Gordon, S., Pattern recognition receptors. Doubling up for the innate immune response. *Cell*, 2002. **111**(7): p. 927-30.
11. Geijtenbeek, T.B.H., et al., Self- and nonself-recognition by C-type lectins on dendritic cells. *Annu Rev Immunol*, 2004. **22**(1): p. 33-54.
12. Taylor, M.E., et al., Primary structure of the mannose receptor contains multiple motifs resembling carbohydrate-recognition domains. *J Biol Chem*, 1990. **265**(21): p. 12156-62.
13. Rosetto, M., et al., Signals from the IL-1 receptor homolog, Toll, can activate an immune response in a *Drosophila* hemocyte cell line. *Biochem. Biophys. Res. Commun.*, 1995. **209**: p. 111-116.
14. Lemaitre, B., et al., The dorsoventral regulatory gene cassette spatzle/toll/cactus controls the potent antifungal response in *Drosophila* adults. *Cell*, 1996. **86**(6): p. 973-83.
15. Gay, N.J. and F.J. Keith, *Drosophila* Toll and IL-1 receptor. *Nature*, 1991. **351**: p. 355-356.
16. Baeurle, P.A. and T. Henkel, Function and activation of NF-kappa B in the immune system. *Annu Rev Immunol*, 1994. **12**: p. 141-79.
17. Levashina, E.A., et al., Constitutive activation of toll-mediated antifungal defense in serpin-deficient *Drosophila*. *Science*, 1999. **285**(5435): p. 1917-9.
18. Ligoxygakis, P., et al., Activation of *Drosophila* Toll during fungal infection by a blood serine protease. *Science*, 2002. **297**(5578): p. 114-116.
19. Takeda, K., T. Kaisho, and S. Akira, Toll-like receptors. *Annu Rev Immunol*, 2003. **21**(1): p. 335-376.

20. Schnare, M., et al., Toll-like receptors control activation of adaptive immune responses. *Nat Immunol*, 2001. **2**(10): p. 947-50.
21. Rosenberg, H.F. and J.I. Gallin, *Inflammation*, in *Fundamental immunology*, W.E. Paul, Editor. 1999, Lippincott-Raven Publishers: Philadelphia. p. 1051-1066.
22. Kushner, I. and A. Mackiewicz, *The acute phase response: An overview*, in *Acute phase proteins : Molecular biology, biochemistry, and clinical applications*, A. Mackiewicz, I. Kushner, and H. Baumann, Editors. 1993, CRC Press: Boca Raton. p. 3-20.
23. Winkelstein, J.A. and A. Tomasz, Activation of the alternative complement pathway by pneumococcal cell wall teichoic acid. *J. Immunol.*, 1978. **120**: p. 174-178.
24. Culley, F.J., et al., C-reactive protein binds to phosphorylated carbohydrates. *Glycobiology*, 2000. **10**(1): p. 59-65.
25. Gewurz, H., X.H. Zhang, and T.F. Lint, Structure and function of the pentraxins. *Curr Opin Immunol*, 1995. **7**(1): p. 54-64.
26. Zweigner, J., et al., High concentrations of lipopolysaccharide-binding protein in serum of patients with severe sepsis or septic shock inhibit the lipopolysaccharide response in human monocytes. *Blood*, 2001. **98**: p. 3800-3808.
27. Wright, S.D., et al., CD14, a receptor for complexes of lipopolysaccharide (lps) and lps binding protein. *Science*, 1990. **249**(4975): p. 1431-3.
28. Triantafilou, M. and K. Triantafilou, Lipopolysaccharide recognition: CD14, TLRs and the LPS-activation cluster. *Trends Immunol*, 2002. **23**(6): p. 301-304.
29. Weber, J.R., et al., Recognition of pneumococcal peptidoglycan. An expanded, pivotal role for LPS binding protein. *Immunity*, 2003. **19**(2): p. 269-79.
30. Dziarski, R., R.I. Tapping, and P.S. Tobias, Binding of bacterial peptidoglycan to CD14. *J Biol Chem*, 1998. **273**(15): p. 8680-90.
31. Drickamer, K., Two distinct classes of carbohydrate-recognition domains in animal lectins. *J Biol Chem*, 1988. **263**(20): p. 9557-9560.
32. Powell, L.D. and A. Varki, I-type lectins. *J Biol Chem*, 1995. **270**(24): p. 14243-14246.
33. Dahms, N.M. and M.K. Hancock, P-type lectins. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 2002. **1572**(2-3): p. 317-340.
34. Coe, J.E., Homologs of crp: A diverse family of proteins with similar structure. *Contemp Top Mol Immunol*, 1983. **9**: p. 211-38.
35. Sharon, N. and H. Lis, Legume lectins-a large family of homologous proteins. *Faseb J*, 1990. **4**(14): p. 3198-208.
36. Sharon, N., Lectin-carbohydrate complexes of plants and animals: An atomic view. *Trends Biochem Sci*, 1993. **18**(6): p. 221-6.
37. Roberts, D.L., et al., Molecular basis of lysosomal enzyme recognition: Three-dimensional structure of the cation-dependent mannose 6-phosphate receptor. *Cell*, 1998. **93**: p. 639-648.
38. May, A.P., et al., Crystal structure of the n-terminal domain of sialoadhesin in complex with 3' sialyllactose at 1.85 Å resolution. *Mol Cell*, 1998. **1**: p. 719-728.
39. Weis, W.I., K. Drickamer, and W.A. Hendrickson, Structure of a C-type mannose-binding protein complexed with an oligosaccharide. *Nature*, 1992. **360**(6400): p. 127-34.
40. Emsley, J., et al., Structure of pentameric human serum amyloid P component. *Nature*, 1994. **367**(6461): p. 338-345.
41. Liao, D.I., et al., Structure of s-lectin, a developmentally regulated vertebrate beta-galactoside-binding protein. *Proc Natl Acad Sci U S A*, 1994. **91**(4): p. 1428-32.

42. Loris, R., et al., Legume lectin structure. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, 1998. **1383**(1): p. 9-36.
43. Kairies, N., et al., The 2.0-Å crystal structure of tachylectin 5a provides evidence for the common origin of the innate immunity and the blood coagulation systems. *Proc Natl Acad Sci U S A*, 2001. **98**(24): p. 13519-24.
44. Beisel, H.-G., et al., Tachylectin-2: Crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the japanese horseshoe crab *Tachypleus tridentatus*. *EMBO J*, 1999. **18**(9): p. 2313-2322.
45. Parodi, A.J., Protein glucosylation and its role in protein folding. *Annu. Rev. Biochem.*, 2000. **69**(1): p. 69-93.
46. Cebo, C., G. Vergoten, and J.-P. Zanetta, Lectin activities of cytokines: Functions and putative carbohydrate-recognition domains. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 2002. **1572**(2-3): p. 422-434.
47. Saxon, E. and C.R. Bertozzi, Chemical and biological strategies for engineering cell surface glycosylation. *Annu Rev Cell Dev Biol*, 2001. **17**(1): p. 1-23.
48. Haines, N. and K.D. Irvine, Glycosylation regulates notch signalling. *Nat Rev Mol Cell Biol*, 2003. **4**(10): p. 786-97.
49. Rudenko, G., E. Hohenester, and Y.A. Muller, LG/LNS domains: Multiple functions - one business end? *Trends Biochem Sci*, 2001. **26**(6): p. 363-8.
50. Drickamer, K., Making a fitting choice: Common aspects of sugar-binding sites in plant and animal lectins. *Structure*, 1997. **5**(4): p. 465-8.
51. Weis, W.I. and K. Drickamer, Structural basis of lectin-carbohydrate recognition. *Annu. Rev. Biochem.*, 1996. **65**: p. 441-73.
52. Drickamer, K., Ca(2+)-dependent sugar recognition by animal lectins. *Biochem Soc Trans*, 1996. **24**(1): p. 146-50.
53. Figdor, C.G., Y. van Kooyk, and G.J. Adema, C-type lectin receptors on dendritic cells and langerhans cells. *Nat Rev Immunol*, 2002. **2**(2): p. 77-84.
54. Vasta, G.R. and J.J. Marchalonis, Humoral and cell membrane-associated lectins from invertebrates and lower chordates: Specificity, molecular characterization and their structural relationships with putative recognition molecules from vertebrates. *Dev Comp Immunol*, 1985. **9**(3): p. 531-9.
55. Vasta, G.R., M.E. Chiesa, and M. Palatnik, Agglutinins and protectins in the snail *Borus*. *Medicina*, 1976. **36**(2): p. 107-12.
56. Vasta, G.R., T.C. Cheng, and J.J. Marchalonis, A lectin on the hemocyte membrane of the oyster (*Crassostrea virginica*). *Cell Immunol*, 1984. **88**(2): p. 475-88.
57. Kotani, E., et al., Cloning and expression of the gene of hemocytin, an insect humoral lectin which is homologous with the mammalian von Willebrand factor. *Biochim Biophys Acta*, 1995. **1260**(3): p. 245-58.
58. Vasta, G.R. and E. Cohen, Humoral lectins in the scorpion *Vaejovis confuscus*: A serological characterization. *J Invertebr Pathol*, 1984. **43**(2): p. 226-33.
59. Vasta, G.R. and E. Cohen, Sialic acid binding lectins in the serum of american spiders of the genus *Aphonopelma*. *Dev Comp Immunol*, 1984. **8**(3): p. 515-22.
60. Takahashi, H., et al., Cloning and sequencing of cDNA of *Sarcophaga peregrina* humoral lectin induced on injury of the body wall. *J Biol Chem*, 1985. **260**(22): p. 12228-33.

61. Cassels, F.J., J.J. Marchalonis, and G.R. Vasta, Heterogeneous humoral and hemocyte-associated lectins with N- acylaminosugar specificities from the blue crab, *Callinectes sapidus* Rathbun. *Comp Biochem Physiol [B]*, 1986. **85**(1): p. 23-30.
62. Vasta, G.R. and E. Cohen, Carbohydrate specificities of *Birgus latro* (coconut crab) serum lectins. *Dev Comp Immunol*, 1984. **8**(1): p. 197-202.
63. Kawabata, S.-i. and S. Iwanaga, Role of lectins in the innate immunity of horseshoe crab. *Dev Comp Immunol*, 1999. **23**(4-5): p. 391-400.
64. Fujita, Y., et al., A novel lectin from sarcophaga. Its purification, characterization, and cDNA cloning. *J Biol Chem*, 1998. **273**(16): p. 9667-72.
65. Haq, S., et al., Purification, characterization, and cDNA cloning of a galactose-specific C-type lectin from *Drosophila melanogaster*. *J Biol Chem*, 1996. **271**(33): p. 20213-20218.
66. Umetsu, K., S. Kosada, and T. Suzuki, Purification and characterization of a lectin from the beetle, *Allomyrina dichotoma*. *J Biochem*, 1984. **95**(1): p. 239-245.
67. Muramoto, K. and H. Kamiya, The amino-acid sequence of multiple lectins of the acorn barnacle *Megabalanus rosa* and its homology with animal lectins. *Biochim Biophys Acta*, 1990. **1039**(1): p. 42-51.
68. Yu, X.Q., H. Gan, and M.R. Kanost, Immulectin, an inducible C-type lectin from an insect, *Manduca sexta*, stimulates activation of plasma prophenol oxidase. *Insect Biochem Mol Biol*, 1999. **29**(7): p. 585-97.
69. Chen, S.-C., et al., Biochemical properties and cDNA cloning of two new lectins from the plasma of *Tachypleus tridentatus*. *Tachypleus* plasma lectin 1 and 2+. *J Biol Chem*, 2001. **276**(13): p. 9631-9639.
70. Wilson, R., C. Chen, and N.A. Ratcliffe, Innate immunity in insects: The role of multiple, endogenous serum lectins in the recognition of foreign invaders in the cockroach, *Blaberus discoidalis*. *J Immunol*, 1999. **162**(3): p. 1590-6.
71. Giga, Y., K. Sutoh, and A. Ikai, A new multimeric hemmagglutinin from the coelomic fluid of the sea urchin *Anthocidaris crassispina*. *Biochemistry*, 1985. **24**: p. 4461-4467.
72. Hatakeyama, T., et al., Amino acid sequence of a C-type lectin Cel-IV from the marine invertebrate *Cucumaria echinata*. *Biosci Biotechnol Biochem*, 1995. **59**(7): p. 1314-7.
73. Snowden, A.M. and G.R. Vasta, A dimeric lectin from coelomic fluid of the starfish *Oreaster reticulatus* cross-reacts with the sea urchin embryonic substrate adhesion protein, echinonectin. *Ann N Y Acad Sci*, 1994. **712**: p. 327-9.
74. Giga, Y., A. Ikai, and K. Takahashi, The complete amino acid sequence of echinoidin, a lectin from the coelomic fluid of the sea urchin *Anthocidaris crassispina*. Homologies with mammalian and insect lectins. *J Biol Chem*, 1987. **262**(13): p. 6197-203.
75. Vasta, G.R., et al., Galactosyl-binding lectins from the tunicate *Didemnum candidum*. Purification and physicochemical characterization. *J Biol Chem*, 1986. **261**(20): p. 9174-81.
76. Rogener, W. and G. Uhlenbruck, Invertebrate lectins: The biological role of a biological rule. *Dev Comp Immunol*, 1984. **Supplement 3**: p. 159-164.
77. Vasta, G.R., et al., Animal lectins as self/non-self recognition molecules. Biochemical and genetic approaches to understanding their biological roles and evolution. *Ann N Y Acad Sci*, 1994. **712**: p. 55-73.
78. Iwanaga, S., The *Limulus* clotting reaction. *Curr Opin Immunol*, 1993. **5**(1): p. 74-82.
79. Levin, J. and F.B. Bang, The role of endotoxin in the extracellular coagulation of *Limulus* blood. *Bull Johns Hopkins Hosp*, 1964. **115**: p. 265-274.

80. Iwanaga, S., S. Kawabata, and T. Muta, New types of clotting factors and defense molecules found in horseshoe crab hemolymph: Their structures and functions. *J Biochem (Tokyo)*, 1998. **123**(1): p. 1-15.
81. The-C.-elegans-Sequencing-Consortium, Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science*, 1998. **282**(5396): p. 2012-2018.
82. Adams, M.D., et al., The genome sequence of *Drosophila melanogaster*. *Science*, 2000. **287**(5461): p. 2185-95.
83. Drickamer, K. and R.B. Dodd, C-type lectin-like domains in *Caenorhabditis elegans*: Predictions from the complete genome sequence. *Glycobiology*, 1999. **9**(12): p. 1357-69.
84. Dodd, R.B. and K. Drickamer, Lectin-like proteins in model organisms: Implications for evolution of carbohydrate-binding activity. *Glycobiology*, 2001. **11**(5): p. 71R-9R.
85. Azumi, K., et al., Genomic analysis of immunity in a urochordate and the emergence of the vertebrate immune system: "waiting for godot". *Immunogenetics*, 2003. **55**(8): p. 570-81.
86. Weis, W.I., M.E. Taylor, and K. Drickamer, The C-type lectin superfamily in the immune system. *Immunol. Rev.*, 1998. **163**: p. 19-34.
87. Brennan, C.A. and K.V. Anderson, *Drosophila*. The genetics of innate immune recognition and response. *Annu Rev Immunol*, 2004. **22**(1): p. 457-483.
88. Matsushita, M., The lectin pathway of the complement system. *Microbiol Immunol*, 1996. **40**(12): p. 887-93.
89. Fujita, T., Evolution of the lectin-complement pathway and its role in innate immunity. *Nat Rev Immunol*, 2002. **2**(5): p. 346-53.
90. Ikeda, K., et al., Serum lectin with known structure activates complement through the classical pathway. *J Biol Chem*, 1987. **262**(16): p. 7451-4.
91. Drickamer, K., M.S. Dordal, and L. Reynolds, Mannose-binding proteins isolated from rat liver contain carbohydrate-recognition domains linked to collagenous tails. Complete primary structures and homology with pulmonary surfactant apoprotein. *J Biol Chem*, 1986. **261**(15): p. 6878-87.
92. Holmskov, U. and J.C. Jensenius, Structure and function of collectins: Humoral C-type lectins with collagenous regions. *Behring Inst Mitt*, 1993(93): p. 224-35.
93. Sheriff, S., C.Y. Chang, and R.A. Ezekowitz, Human mannose-binding protein carbohydrate recognition domain trimerizes through a triple alpha-helical coiled-coil. *Nat Struct Biol*, 1994. **1**(11): p. 789-94.
94. Wallis, R. and K. Drickamer, Molecular determinants of oligomer formation and complement fixation in mannose-binding proteins. *J Biol Chem*, 1999. **274**(6): p. 3580-9.
95. Wallis, R. and R.B. Dodd, Interaction of mannose-binding protein with associated serine proteases. Effects of naturally occurring mutations. *J Biol Chem*, 2000. **275**(40): p. 30962-30969.
96. Lu, J. and Y. Le, Ficolins and the fibrinogen-like domain. *Immunobiology*, 1998. **199**(2): p. 190-9.
97. Hansen, S. and U. Holmskov, Structural aspects of collectins and receptors for collectins. *Immunobiology*, 1998. **199**(2): p. 165-89.
98. Matsushita, M. and T. Fujita, Activation of the classical complement pathway by mannose-binding protein in association with a novel c1s-like serine protease. *J. Exp. Med.*, 1992. **176**: p. 1497.

99. Agrawal, A., et al., Topology and structure of the c1q-binding site on C-reactive protein. *J Immunol*, 2001. **166**(6): p. 3998-4004.
100. Sun, S.C., et al., Hemolin: An insect-immune protein belonging to the immunoglobulin superfamily. *Science*, 1990. **250**(4988): p. 1729-32.
101. Yu, X.Q. and M.R. Kanost, Binding of hemolin to bacterial lipopolysaccharide and lipoteichoic acid. An immunoglobulin superfamily member from insects as a pattern-recognition receptor. *Eur J Biochem*, 2002. **269**(7): p. 1827-34.
102. Schluter, S.F., et al., 'Big bang' emergence of the combinatorial immune system. *Dev Comp Immunol*, 1999. **23**(2): p. 107-11.
103. Kasahara, M., What do the paralogous regions in the genome tell us about the origin of the adaptive immune system? *Immunol. Rev.*, 1998. **166**: p. 159-75.
104. Kasahara, M., T. Suzuki, and L.D. Pasquier, On the origins of the adaptive immune system: Novel insights from invertebrates and cold-blooded vertebrates. *Trends in Immunology*, 2004. **25**(2): p. 105-111.
105. Ohno, S., Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. *Semin Cell Dev Biol*, 1999. **10**(5): p. 517-22.
106. Chothia, C., et al., Evolution of the protein repertoire. *Science*, 2003. **300**(5626): p. 1701-1703.
107. Spring, J., Vertebrate evolution by interspecific hybridisation--are we polyploid? *FEBS Lett*, 1997. **400**(1): p. 2-8.
108. McLysaght, A., K. Hokamp, and K.H. Wolfe, Extensive genomic duplication during early chordate evolution. *Nat Genet*, 2002. **31**(2): p. 200-4.
109. Vandepoele, K., et al., Major events in the genome evolution of vertebrates: Paraneome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci U S A*, 2004: p. 0307968100.
110. Taylor, J.S., et al., Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res*, 2003. **13**(3): p. 382-390.
111. Warr, G.W., The immunoglobulin genes of fish. *Dev Comp Immunol*, 1995. **19**(1): p. 1-12.
112. Du Pasquier, L., The immune system of invertebrates and vertebrates. *Comp Biochem Physiol B Biochem Mol Biol*, 2001. **129**(1): p. 1-15.
113. Wienholds, E., et al., Target-selected inactivation of the zebrafish rag1 gene. *Science*, 2002. **297**(5578): p. 99-102.
114. Meijer, A.H., et al., Expression analysis of the Toll-like receptor and TIR domain adaptor families of zebrafish. *Mol. Immunol.*, 2004. **40**(11): p. 773-783.
115. Oshiumi, H., et al., Prediction of the prototype of the human Toll-like receptor gene family from the pufferfish, *Fugu rubripes*, genome. *Immunogenetics*, 2003. **54**(11): p. 791-800.
116. Jault, C., L. Pichon, and J. Chluba, Toll-like receptor gene family and TIR-domain adaptors in *Danio rerio**1. *Mol. Immunol.*, 2004. **40**(11): p. 759-771.
117. Sunyer, J.O., I.K. Zarkadis, and J.D. Lambris, Complement diversity: A mechanism for generating immune diversity? *Immunol Today*, 1998. **19**(11): p. 519-23.
118. Holmskov, U., S. Thiel, and J.C. Jensenius, Collectins and ficolins: Humoral lectins of the innate immune defense. *Annu Rev Immunol*, 2003. **21**(1): p. 547-578.
119. Cartwright, J.R., et al., Isolation and characterisation of pentraxin-like serum proteins from the common carp *Cyprinus carpio*. *Dev Comp Immunol*, 2004. **28**(2): p. 113-125.

120. Lund, V. and J.A. Olafsen, A comparative study of pentraxin-like proteins in different fish species. *Dev Comp Immunol*, 1998. **22**(2): p. 185-94.
121. Cook, M.T., et al., Isolation and partial characterization of a pentraxin-like protein with complement-fixing activity from snapper (*Pagrus auratus*, Sparidae) serum. *Dev Comp Immunol*, 2003. **27**(6-7): p. 579-88.
122. Jensen, L.E., et al., Acute phase response in salmonids: Evolutionary analyses and acute phase response. *J Immunol*, 1997. **158**(1): p. 384-92.
123. Inagawa, H., et al., Cloning and characterization of the homolog of mammalian lipopolysaccharide-binding protein and bactericidal permeability-increasing protein in rainbow trout *Oncorhynchus mykiss*. *J Immunol*, 2002. **168**(11): p. 5638-5644.
124. Nakao, M., et al., Molecular cloning of the complement (c1r/c1s/masp2-like serine proteases from the common carp (*Cyprinus carpio*). *Immunogenetics*, 2001. **52**(3-4): p. 255-63.
125. Vitved, L., et al., The homologue of mannose-binding lectin in the carp family cyprinidae is expressed at high level in spleen, and the deduced primary structure predicts affinity for galactose. *Immunogenetics*, 2000. **51**(11): p. 955-64.
126. Manihar, S.R. and H.R. Das, Isolation and characterization of a new lectin from plasma of fish channa punctatus. *Biochimica et Biophysica Acta*, 1990. **1036**: p. 162-165.
127. Jensen, L.E., et al., A rainbow trout lectin with multimeric structure. *Comp Biochem Physiol B Biochem Mol Biol*, 1997. **116**(4): p. 385-390.
128. Tateno, H., et al., A novel rhamnose-binding lectin family from eggs of steelhead trout (*Oncorhynchus mykiss*) with different structures and tissue distribution. *Biosci Biotechnol Biochem*, 2001. **65**(6): p. 1328-38.
129. Hoover, G.J., et al., Plasma proteins of rainbow trout (*Oncorhynchus mykiss*) isolated by binding to lipopolysaccharide from aeromonas salmonicida. *Comp Biochem Physiol B Biochem Mol Biol*, 1998. **120**(3): p. 559-69.
130. Miller, N., et al., Functional and molecular characterization of teleost leukocytes. *Immunol. Rev.*, 1998. **166**: p. 187-97.
131. Ganassin, R.C. and N.C. Bols, Development of a monocyte/macrophage-like cell line, RTS11, from rainbow trout spleen. *Fish & Shellfish Immunology*, 1998. **8**(6): p. 457-476.
132. Secombes, C.J., et al., Cytokine genes in fish. *Aquaculture*, 1999. **172**(1-2): p. 93-102.
133. Yoder, J.A., et al., Zebrafish as an immunological model system. *Microbes and Infection*, 2002. **4**(14): p. 1469-1478.
134. Evelyn, T.P., A historical review of fish vaccinology. *Dev Biol Stand*, 1997. **90**: p. 3-12.
135. Leclerc, G.M., et al., Characterization of a highly repetitive sequence conserved among the north american *Morone* species. *Marine Biotechnology*, 1999. **1**(2): p. 122-130.
136. Whitehurst, D.K. and R.E. Stevens, *History and overview of striped bass culture and management*, in *Culture and propagation of striped bass and its hybrids*, R.M. Harrell, J.H. Kerby, and R.V. Minton, Editors. 1990, American Fisheries Society: Bethesda, Maryland. p. 1-5.
137. Kerby, J.H. and R.M. Harrell, *Hybridization, genetic manipulation, and gene pool conservation of striped bass*, in *Culture and propagation of striped bass and its hybrids*, R.M. Harrell, J.H. Kerby, and R.V. Minton, Editors. 1990, American Fisheries Society: Bethesda, Maryland. p. 159-190.
138. Noga, E.J., et al., Acute stress causes skin ulceration in striped bass and hybrid bass (*Morone*). *Vet Pathol*, 1998. **35**(2): p. 102-7.

139. Reubush, K.J. and A.G. Heath, Secondary stress responses to acute handling in striped bass (*morone saxatilis*) and hybrid striped bass (*Morone chrysops* \times *Morone saxatilis*). *Am J Vet Res*, 1997. **58**(12): p. 1451-6.
140. Harrell, R.M., ed. *Striped bass and other morone culture*. Developments in aquaculture and fisheries science. Vol. 30. 1997, Elsevier. 386.
141. Franc, N.C., et al., Croquemort, a novel *Drosophila* hemocyte/macrophage receptor that recognizes apoptotic cells. *Immunity*, 1996. **4**(5): p. 431-43.
142. Kawasaki, T., R. Etoh, and I. Yamashina, Isolation and characterization of a mannan-binding protein from rabbit liver. *Biochemical and Biophysical Research Communications*, 1978. **81**(3): p. 1018-1024.
143. Mizuno, Y., et al., Isolation and characterization of a mannan-binding protein from rat liver. *J Biol Chem*, 1981. **256**(9): p. 4247-4252.
144. Lee, R.T., et al., Major lectin of alligator liver is specific for mannose/L-fucose. *J Biol Chem*, 1994. **269**(30): p. 19617-25.
145. Kurata, H., et al., Structure and function of mannan-binding proteins isolated from human liver and serum. *J Biochem (Tokyo)*, 1994. **115**(6): p. 1148-54.
146. Ahmed, H., et al., Galectin-1 from bovine spleen: Biochemical characterization, carbohydrate specificity and tissue-specific isoform profiles. *J Biochem (Tokyo)*, 1996. **120**(5): p. 1007-19.
147. Bradford, M.M., A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem*, 1976. **72**: p. 248-54.
148. Hermanson, G.T., A.K. Mallia, and P.K. Smith, *Immobilized affinity ligand techniques*. 1992, San Diego: Academic Press. xxiii, 454.
149. Kawasaki, T. and G. Ashwell, Isolation and characterization of an avian hepatic binding protein specific for n-acetylglucosamide-terminated glycoproteins. *J Biol Chem*, 1977. **252**(18): p. 6536-6543.
150. Laemmli, U.K., Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, 1970. **227**(259): p. 680-5.
151. Coligan, J.E., et al., eds. *Current protocols in protein science*. 1995, Wiley: Brooklyn, N.Y. v. (looseleaf).
152. Vasta, G.R. and J.J. Marchalonis, Galactosyl-binding lectins from the tunicate *Didemnum candidum*. Carbohydrate specificity and characterization of the combining site. *J Biol Chem*, 1986. **261**(20): p. 9182-6.
153. Hewick, R.M., et al., A gas-liquid solid phase peptide and protein sequenator. *J Biol Chem*, 1981. **256**(15): p. 7990-7.
154. Pohl, J., Sequence analysis of peptide resins from BOC/benzyl solid-phase synthesis. *Methods Mol Biol*, 1994. **36**: p. 107-29.
155. Chomczynski, P. and N. Sacchi, Single-step method of rna isolation by acid guanidinium thiocyanate- phenol-chloroform extraction. *Anal Biochem*, 1987. **162**(1): p. 156-9.
156. Green, E.D., *Genome analysis : A laboratory manual*. 1997, Plainview, N.Y.: Cold Spring Harbor Laboratory Press. 4 v.
157. Ausubel, F.M., et al., *Current protocols in molecular biology*. 1988, New York: Published by Greene Pub. Associates and Wiley- Interscience : J. Wiley. v. (looseleaf).

158. Sambrook, J., T. Maniatis, and E.F. Fritsch, *Molecular cloning : A laboratory manual*. 2nd ed. 1989, Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory. 3 v.
159. Lee, C.C. and C.T. Caskey, *cDNA cloning using degenerate primers*. PCR protocols : A guide to methods and applications, ed. M.A. Innis. 1990, San Diego: Academic Press. xvii,482.
160. Compton, T., *Degenerate primers for DNA amplification*, in *PCR protocols : A guide to methods and applications*, M.A. Innis, Editor. 1990, Academic Press: San Diego. p. xvii,482.
161. Davis, L.G., W.M. Kuehl, and J.F. Battey, *Basic methods in molecular biology*. 2nd ed. 1994, Norwalk, Conn.: Appleton & Lange. xiv, 777.
162. Nielsen, H., et al., Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, 1997. **10**: p. 1-6.
163. Lee, R.T., et al., Ligand-binding characteristics of rat serum-type mannose-binding protein (mbp-a). Homology of binding site architecture with mammalian and chicken hepatic lectins. *J Biol Chem*, 1991. **266**(8): p. 4810-5.
164. Townsend, R. and P. Stahl, Isolation and characterization of a mannose/n-acetylglucosamine/fucose- binding protein from rat liver. *Biochem J*, 1981. **194**(1): p. 209-14.
165. Baenziger, J. and Y. Maynard, Human hepatic lectin. Physicochemical properties and specificity. *J Biol Chem*, 1980. **255**(10): p. 4607-4613.
166. Pilstrom, L. and E. Bengten, Immunoglobulin in fish-genes, expression and structure. *Fish & Shellfish Immunology*, 1996. **6**(4): p. 243-262.
167. Flajnik, M.F. and L.L. Rumfelt, Early and natural antibodies in non-mammalian vertebrates. *Curr Top Microbiol Immunol*, 2000. **252**: p. 233-40.
168. Kelly, C., Physicochemical properties and n-terminal sequence of eel lectin. *Biochemical Journal*, 1984. **220**(1): p. 221-226.
169. Gercken, J. and L. Renwranzt, A new mannan-binding lectin from the serum of the eel (*Anguilla anguilla* l.): Isolation, characterization and comparison with the fucose-specific serum lectin. *Comparative Biochemistry & Physiology Biochemistry & Molecular Biology*, 1994. **108**(4): p. 449-61.
170. Wiese, T., J. Dunlap, and M. Yorek, L-fucose is accumulated via a specific transport system in eukaryotic cells. *J Biol Chem*, 1994. **269**(36): p. 22705-22711.
171. Flowers, H.M., Chemistry and biochemistry of D- and L-fucose. *Adv Carbohydr Chem Biochem*, 1981. **39**: p. 279-345.
172. Ng, K.K., K. Drickamer, and W.I. Weis, Structural analysis of monosaccharide recognition by rat liver mannose-binding protein. *J Biol Chem*, 1996. **271**(2): p. 663-74.
173. D'Arcy, S.M., et al., Determination of the structure of a novel acidic oligosaccharide with blood-group activity isolated from bovine submaxillary-gland mucin. *Biochem J*, 1989. **260**(2): p. 389-93.
174. Slomiany, B.L. and K. Meyer, Oligosaccharides produced by acetolysis of blood group active (A + H)sulfated glycoproteins from hog gastric mucin. *J. Biol. Chem.*, 1973. **248**(7): p. 2290-2295.
175. Patankar, M.S., et al., A revised structure for fucoidan may explain some of its biological activities. *J Biol Chem*, 1993. **268**(29): p. 21770-6.
176. Mitchell, E., et al., Structural basis for oligosaccharide-mediated adhesion of *Pseudomonas aeruginosa* in the lungs of cystic fibrosis patients. *Nat Struct Biol*, 2002. **9**(12): p. 918-21.

177. Loris, R., et al., Structural basis of carbohydrate recognition by the lectin LecB from *Pseudomonas aeruginosa*. *J Mol Biol*, 2003. **331**(4): p. 861-870.
178. Audette, G.F., M. Vandonseelaar, and L.T. Delbaere, The 2.2 Å resolution structure of the O(H) blood-group-specific lectin I from *Ulex europaeus* Dagger. *J Mol Biol*, 2000. **304**(3): p. 423-33.
179. Wimmerova, M., et al., Crystal structure of fungal lectin: Six-bladed beta-propeller fold and novel fucose recognition mode for *Aleuria aurantia* lectin. *J Biol Chem*, 2003. **278**(29): p. 27059-27067.
180. Fujihashi, M., et al., Crystal structure of fucose-specific lectin from *Aleuria aurantia* binding ligands at three of its five sugar recognition sites. *Biochemistry*, 2003. **42**(38): p. 11093-11099.
181. Saito, T., et al., A newly identified horseshoe crab lectin with binding specificity to O-antigen of bacterial lipopolysaccharides. *J Biol Chem*, 1997. **272**(49): p. 30703-8.
182. Honda, S., et al., Multiplicity, structures, and endocrine and exocrine natures of eel fucose-binding lectins. *J Biol Chem*, 2000. **275**(42): p. 33151-7.
183. Watkins, W.M. and W.T.J. Morgan, Neutralisation of the anti-H agglutinin in eel serum by simple sugars. *Nature*, 1952. **169**: p. 825-826.
184. Yano, T., *The nonspecific immune system: Humoral defenses*, in *The fish immune system: Organism, pathogen, and environment*, G. Iwama and T. Nakanishi, Editors. 1996, Academic Press. p. 105-157.
185. Kalb, A.J., et al., Manganese(II) in concanavalin a and other lectin proteins. *Met Ions Biol Syst*, 2000. **37**: p. 279-304.
186. Ezekowitz, R.A., et al., Molecular characterization of the human macrophage mannose receptor: Demonstration of multiple carbohydrate recognition-like domains and phagocytosis of yeasts in cos-1 cells. *J Exp Med*, 1990. **172**(6): p. 1785-94.
187. Oda, Y., et al., Soluble lactose-binding lectin from rat intestine with two different carbohydrate-binding domains in the same peptide chain. *J Biol Chem*, 1993. **268**(8): p. 5929-39.
188. Hadari, Y.R., et al., Galectin-8. A new rat lectin, related to galectin-4. *J Biol Chem*, 1995. **270**(7): p. 3447-53.
189. Saito, T., et al., A novel type of limulus lectin-l6: Purification, primary structure, and antibacterial activity. *J Biol Chem*, 1995. **270**(24): p. 14493-14499.
190. Okino, N., et al., Purification, characterization, and cDNA cloning of a 27-Kda lectin (L10) from horseshoe crab hemocytes. *J Biol Chem*, 1995. **270**(52): p. 31008-31015.
191. Quesenberry, M., et al., Tunicate *Clavelina picta* lectin III contains carbohydrate-recognition domains homologous to vertebrate serum mannose-binding proteins. *Glycobiology*, 1996. **6**(7): p. 411-411.
192. Tateno, H., et al., Isolation and characterization of rhamnose-binding lectins from eggs of steelhead trout (*Oncorhynchus mykiss*) homologous to low density lipoprotein receptor superfamily. *J Biol Chem*, 1998. **273**(30): p. 19190-7.
193. Kilpatrick, D.C., Animal lectins: A historical introduction and overview. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 2002. **1572**(2-3): p. 187-197.
194. Koizumi, N., et al., The lipopolysaccharide-binding protein participating in hemocyte nodule formation in the silkworm *Bombyx mori* is a novel member of the C-type lectin superfamily with two different tandem carbohydrate-recognition domains. *FEBS Lett*, 1999. **443**(2): p. 139-43.

195. Glibetic, M.D. and H. Baumann, Influence of chronic inflammation on the level of mRNA for acute-phase reactants in the mouse liver. *J Immunol*, 1986. **137**(5): p. 1616-22.
196. Lin, L. and T.Y. Liu, Isolation and characterization of c-reactive protein (CRP) cDNA and genomic DNA from xenopus laevis. A species representing an intermediate stage in crp evolution. *J Biol Chem*, 1993. **268**(9): p. 6809-15.
197. Borson, N.D., W.L. Salo, and L.R. Drewes, A lock-docking oligo(dT) primer for 5' and 3' RACE PCR. *PCR Methods Appl*, 1992. **2**(2): p. 144-8.
198. Dieffenbach, C.W. and G.S. Dveksler, *Pcr primer: A laboratory manual*. 1995, Plainview, N.Y.: Cold Spring Harbor Laboratory Press. xii, 714.
199. Altschul, S.F., et al., Basic local alignment search tool. *J Mol Biol*, 1990. **215**: p. 403-10.
200. Boguski, M.S., T.M. Lowe, and C.M. Tolstoshev, dbEST--database for "expressed sequence tags". *Nat Genet*, 1993. **4**(4): p. 332-3.
201. Rose, T.M., et al., Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Res*, 1998. **26**(7): p. 1628-35.
202. O'Connell, P.O. and M. Rosbash, Sequence, structure, and codon preference of the *Drosophila* ribosomal protein 49 gene. *Nucleic Acids Res*, 1984. **12**(13): p. 5495-513.
203. Thompson, J.D., et al., The Clustal_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*, 1997. **25**(24): p. 4876-82.
204. Nicholas, K.B. and Nicholas H.B. Jr., Genedoc: Analysis and visualization of genetic variation. 1997.
205. Henikoff, S. and J.G. Henikoff, Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 1992. **89**(22): p. 10915-9.
206. Saitou, N. and M. Nei, The neighborjoining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 1987. **4**: p. 406-25.
207. Felsenstein, J., Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 1985. **39**: p. 783-91.
208. Gribskov, M., A.D. McLachlan, and D. Eisenberg, Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci U S A*, 1987. **84**(13): p. 4355-8.
209. Seery, L.T., et al., Identification of a novel member of the pentraxin family in *Xenopus laevis*. *Proceedings of the Royal Society of London-Series B: Biological Sciences*, 1993. **253**(1338): p. 263-270.
210. Tennent, G.A. and M.B. Pepys, Glycobiology of the pentraxins. *Biochem Soc Trans*, 1994. **22**(1): p. 74-9.
211. Lee, G.W., L.T. H., and J. Vilcek, TSG-14, a tumor necrosis factor- and IL-1-inducible protein, is a novel member of the pentraxin family of acute phase response. *J Immunol*, 1993. **150**(5): p. 1804-1812.
212. Gilges, D., et al., Polydom: A secreted protein with pentraxin, complement control protein, epidermal growth factor and von Willebrand factor a domains. *Biochem J*, 2000. **352**(Pt 1): p. 49-59.
213. Goodman, A.R., et al., Long pentraxins: An emerging group of proteins with diverse functions. *Cytokine Growth Factor Rev*, 1996. **7**(2): p. 191-202.
214. Introna, M., et al., Cloning of mouse PTX3, a new member of the pentraxin gene family expressed at extrahepatic sites. *Blood*, 1996. **87**(5): p. 1862-1872.

215. Bottazzi, B., et al., Multimer formation and ligand recognition by the long pentraxin ptx3. *J Biol Chem*, 1997. **272**(52): p. 32817-32823.
216. Okubo, K., et al., Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression *Nat Genet*, 1992. **2**(3): p. 173-9.
217. Adams, M.D., et al., Sequence identification of 2,375 human brain genes. *Nature*, 1992. **355**(6361): p. 632-4.
218. Garrett, J.E., D.S. Knutzon, and D. Carroll, Composite transposable elements in the *Xenopus laevis* genome. *Mol Cell Biol*, 1989. **9**(7): p. 3018-27.
219. Wolfsberg, T.G. and D. Landsman, A comparison of expressed sequence tags (ests) to human genomic sequences. *Nucleic Acids Res*, 1997. **25**(8): p. 1626-32.
220. Hillier, L.D., et al., Generation and analysis of 280,000 human expressed sequence tags. *Genome Res*, 1996. **6**(9): p. 807-28.
221. Alonso, E., et al., Lectin histochemistry shows fucosylated glycoconjugates in the primordial germ cells of *Xenopus* embryos. *J Histochem Cytochem*, 2003. **51**(2): p. 239-243.
222. Beck, C.W. and J.M. Slack, An amphibian with ambition: A new role for *Xenopus* in the 21st century. *Genome Biol*, 2001. **2**(10): p. Reviews1029.
223. Lee, R.T. and Y.C. Lee, Affinity enhancement by multivalent lectin-carbohydrate interaction. *Glycoconj J*, 2000. **17**(7-9): p. 543-51.
224. Duellman, W.E. and L. Trueb, *Biology of amphibians*. 1994, Baltimore: Johns Hopkins University Press. xxi, 670 p.
225. Zeng, S. and Z. Gong, Expressed sequence tag analysis of expression profiles of zebrafish testis and ovary. *Gene*, 2002. **294**(1-2): p. 45-53.
226. Pate, J.L. and P. Landis Keyes, Immune cells in the corpus luteum: Friends or foes? *Reproduction*, 2001. **122**(5): p. 665-76.
227. Jegou, B., The sertoli cell in vivo and in vitro. *Cell Biol Toxicol*, 1992. **8**(3): p. 49-54.
228. Topliss, J.A. and D.J. Rogers, An anti-fucose agglutinin in the ova of *Dicentrarchus labrax*. *Med Lab Sci*, 1985. **42**(2): p. 199-200.
229. Rexroad, C.E., et al., Eighteen polymorphic microsatellite markers for rainbow trout (*Oncorhynchus mykiss*). *Anim Genet*, 2002. **33**(1): p. 76-8.
230. Murai, T., et al., Isolation and characterization of rainbow trout C-reactive protein. *Dev Comp Immunol*, 1990. **14**(1): p. 49-58.
231. Savan, R. and M. Sakai, Analysis of expressed sequence tags (EST) obtained from common carp, cyprinus carpio l., head kidney cells after stimulation by two mitogens, lipopolysaccharide and concanavalin-A. *Comp Biochem Physiol B Biochem Mol Biol*, 2002. **131**(1): p. 71-82.
232. Sagerstrom, C.G., B.I. Sun, and H.L. Sive, Subtractive cloning: Past, present and future. *Annu. Rev. Biochem.*, 1997. **66**: p. 751-83.
233. Ostrander, G.K., *The laboratory fish*. Handbook of experimental animals. 2000, San Diego: Academic Press. xvi, 678 p.
234. McKinnon, J.S. and H.D. Rundle, Speciation in nature: The threespine stickleback model systems. *Trends Ecol Evol*, 2002. **17**(10): p. 480-488.
235. Peichel, C.L., et al., The genetic architecture of divergence between threespine stickleback species. *Nature*, 2001. **414**(6866): p. 901-5.
236. Uhlenbruck, G., E. Janssen, and S. Javeri, Two different anti-galactan lectins in eel serum. *Immunobiology*, 1982. **163**(1): p. 36-47.

237. Aparicio, S., et al., Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, 2002. **297**(5585): p. 1301-1310.
238. Helfman, G.S., B.B. Collette, and D.E. Facey, *The diversity of fishes*. 1997, Malden, Mass.: Blackwell Science. xii, 528.
239. Brainerd, E.L., et al., Patterns of genome size evolution in Tetraodontiform fishes. *Evolution*, 2001. **55**(11): p. 2363-2368.
240. Hoffmann, J.A., et al., Phylogenetic perspectives in innate immunity. *Science*, 1999. **284**(5418): p. 1313-8.
241. Medzhitov, R., P. Preston-Hurlburt, and C.A. Janeway, Jr., A human homologue of the *Drosophila* toll protein signals activation of adaptive immunity. *Nature*, 1997. **388**(6640): p. 394-7.
242. Medzhitov, R., Toll-like receptors and innate immunity. *Nat Rev Immunol*, 2001. **1**(2): p. 135-45.
243. Bentley, D.R., Structural superfamilies of the complement system. *Exp Clin Immunogenet*, 1988. **5**(2-3): p. 69-80.
244. Perez-Vilar, J. and R.L. Hill, The structure and assembly of secreted mucins. *J Biol Chem*, 1999. **274**(45): p. 31751-4.
245. von Heijne, G., Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol*, 1992. **225**(2): p. 487-94.
246. The flybase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res*, 2003. **31**(1): p. 172-5.
247. Leshko-Lindsay, L. and V.G. Corces, The role of selectins in *Drosophila* eye and bristle development. *Development*, 1997. **124**: p. 169-180.
248. Mollereau, B., et al., A green fluorescent protein enhancer trap screen in *Drosophila* photoreceptor cells. *Mech. Dev.*, 2000. **93**(1-2): p. 151-160.
249. Ramet, M., et al., Functional genomic analysis of phagocytosis and identification of a *Drosophila* receptor for *E. coli*. *Nature*, 2002. **416**(6881): p. 644-8.
250. Choe, K.-M., et al., Requirement for a peptidoglycan recognition protein (pgrp) in relish activation and antibacterial immune responses in *Drosophila*. *Science*, 2002. **296**(5566): p. 359-362.
251. Leulier, F., et al., The *Drosophila* immune system detects bacteria through specific peptidoglycan recognition. *Nat Immunol*, 2003. **4**(5): p. 478-84.
252. Wilson, I.B.H., Glycosylation of proteins in plants and invertebrates. *Curr Opin Struct Biol*, 2002. **12**(5): p. 569-577.
253. Weis, W.I., et al., Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by mad phasing. *Science*, 1991. **254**(5038): p. 1608-15.
254. Drickamer, K., C-type lectin-like domains. *Curr Opin Struct Biol*, 1999. **9**(5): p. 585-590.
255. Drickamer, K., Engineering galactose-binding activity into a C-type mannose-binding protein. *Nature*, 1992. **360**(6400): p. 183-6.
256. Natarajan, K., et al., Structure and function of natural killer cell receptors: Multiple molecular solutions to self, nonself discrimination. *Annu Rev Immunol*, 2002. **20**(1): p. 853-885.
257. Holt, R.A., et al., The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, 2002. **298**(5591): p. 129-149.

258. Zdobnov, E.M., et al., Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science*, 2002. **298**(5591): p. 149-159.
259. Kaufman, T.C., D.W. Severson, and G.E. Robinson, The *Anopheles* genome and comparative insect genomics. *Science*, 2002. **298**(5591): p. 97-98.
260. Stormer, L., Phylogeny and taxonomy of fossil horseshoe crab. *J Paleontol*, 1952. **26**: p. 630-639.
261. Parkinson, J., et al., 400,000 nematode ESTs on the net. *Trends Parasitol*, 2003. **19**(7): p. 283-6.
262. Aguinaldo, A.M., et al., Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 1997. **387**(6632): p. 489-93.
263. Adoutte, A., et al., Animal evolution: The end of the intermediate taxa? *Trends Genet*, 1999. **15**(3): p. 104-108.
264. Bernal, A., U. Ear, and N. Kyrpides, Genomes online database (gold): A monitor of genome projects world-wide. *Nucleic Acids Res*, 2001. **29**(1): p. 126-7.
265. King, N., C.T. Hittinger, and S.B. Carroll, Evolution of key cell signaling and adhesion protein families predates animal origins. *Science*, 2003. **301**(5631): p. 361-363.
266. Tettelin, H., et al., Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science*, 2001. **293**(5529): p. 498-506.
267. Avery, O.T., C. MacLeod, and M. McCarty, Studies on the chemical nature of the substance inducing transformation of the pneumococcal types. *J Exp Med*, 1944. **79**: p. 137-158.
268. Peterson, J.D., et al., The comprehensive microbial resource. *Nucl Acids Res*, 2001. **29**(1): p. 123-125.
269. Kim, M.-S., et al., Crystal structures of rbsd leading to the identification of cytoplasmic sugar-binding proteins with a novel folding architecture. *J Biol Chem*, 2003. **278**(30): p. 28173-28180.
270. Di Guilmi, A.M. and A. Dessen, New approaches towards the identification of antibiotic and vaccine targets in *Streptococcus pneumoniae*. *EMBO Reports*, 2002. **3**(8): p. 728-734.
271. Chen, Y.M., Y. Zhu, and E.C. Lin, The organization of the fuc regulon specifying L-fucose dissimilation in *Escherichia coli* k12 as determined by gene cloning. *Mol Gen Genet*, 1987. **210**(2): p. 331-7.
272. Bergey, D.H. and J.G. Holt, *Bergey's manual of determinative bacteriology*. 9th ed. 1994, Baltimore: Williams & Wilkins. xviii, 787.
273. Chan, P.F., et al., Characterization of a novel fucose-regulated promoter (pfcsk) suitable for gene essentiality and antibacterial mode-of-action studies in *Streptococcus pneumoniae*. *J Bacteriol*, 2003. **185**(6): p. 2051-2058.
274. Becker, D.J. and J.B. Lowe, Fucose: Biosynthesis and biological function in mammals. *Glycobiology*, 2003. **13**(7): p. 41R-53.
275. Hoskins, J., et al., Genome of the bacterium *Streptococcus pneumoniae* strain Rr6. *J Bacteriol*, 2001. **183**(19): p. 5709-5717.
276. Shimizu, T., et al., Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater. *Proc Natl Acad Sci U S A*, 2002. **99**(2): p. 996-1001.
277. Boraston, A.B., et al., Co-operative binding of triplicate carbohydrate-binding modules from a thermophilic xylanase. *Mol Microbiol*, 2002. **43**(1): p. 187-94.

278. Servant, F., et al., Prodom: Automated clustering of homologous domains. *Brief Bioinform*, 2002. **3**(3): p. 246-51.
279. Gonzalez, J.M. and R.M. Weiner, Phylogenetic characterization of marine bacterium strain 2-40, a degrader of complex polysaccharides. *Int J Syst Evol Microbiol*, 2000. **50 Pt 2**: p. 831-4.
280. Ensor, L., S. Stosz, and R. Weiner, Expression of multiple complex polysaccharide-degrading enzyme systems by marine bacterium strain 2-40. *J Ind Microbiol Biotechnol*, 1999. **23**(2): p. 123-126.
281. Gaskell, A., S. Crennell, and G. Taylor, The three domains of a bacterial sialidase: A beta-propeller, an immunoglobulin module and a galactose-binding jelly-roll. *Structure*, 1995. **3**(11): p. 1197-205.
282. Ito, N., et al., Novel thioether bond revealed by a 1.7 Å crystal structure of galactose oxidase. *Nature*, 1991. **350**(6313): p. 87-90.
283. Doi, R.H., et al., Cellulosomes from mesophilic bacteria. *J. Bacteriol.*, 2003. **185**(20): p. 5907-5914.
284. Perrin, R.M., et al., Analysis of xyloglucan fucosylation in *Arabidopsis*. *Plant Physiol.*, 2003. **132**(2): p. 768-778.
285. Koonin, E.V., K.S. Makarova, and L. Aravind, Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu Rev Microbiol*, 2001. **55**(1): p. 709-742.
286. Blanchard, J.L. and M. Lynch, Organellar genes: Why do they end up in the nucleus? *Trends Genet*, 2000. **16**(7): p. 315-20.
287. International Human Genome Sequencing Consortium, et al., Initial sequencing and analysis of the human genome. *Nature*, 2001. **409**(6822): p. 860-921.
288. Salzberg, S.L., et al., Microbial genes in the human genome: Lateral transfer or gene loss? *Science*, 2001. **292**(5523): p. 1903-1906.
289. Roelofs, J. and P.J. Van Haastert, Genes lost during evolution. *Nature*, 2001. **411**(6841): p. 1013-4.
290. Stanhope, M.J., et al., Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature*, 2001. **411**(6840): p. 940-4.
291. Bork, P. and R.F. Doolittle, Proposed acquisition of an animal protein domain by bacteria. *Proc Natl Acad Sci U S A*, 1992. **89**(19): p. 8990-4.
292. Hegyi, H., et al., Structural genomics analysis: Characteristics of atypical, common, and horizontally transferred folds. *Proteins*, 2002. **47**(2): p. 126-41.
293. Ponting, C.P., et al., Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J Mol Biol*, 1999. **289**(4): p. 729-45.
294. Yother, J., et al., Genetics of streptococci, lactococci, and enterococci: Review of the sixth international conference. *J Bacteriol*, 2002. **184**(22): p. 6085-6092.
295. Ferretti, J.J., et al., Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc Natl Acad Sci U S A*, 2001. **98**(8): p. 4658-4663.
296. Tettelin, H., et al., Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. *Proc Natl Acad Sci U S A*, 2002. **99**(19): p. 12391-6.
297. Ajdic', D., et al., Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen. *Proc Natl Acad Sci U S A*, 2002. **99**(22): p. 14434-14439.

298. Facklam, R., What happened to the streptococci: Overview of taxonomic and nomenclature changes. *Clin Microbiol Rev*, 2002. **15**(4): p. 613-630.
299. Ochman, H. and N.A. Moran, Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis. *Science*, 2001. **292**(5519): p. 1096-1099.
300. Apic, G., J. Gough, and S.A. Teichmann, An insight into domain combinations. *Bioinformatics*, 2001. **17**(Suppl 1): p. S83-9.
301. Bork, P., et al., Structure and distribution of modules in extracellular proteins. *Q Rev Biophys*, 1996. **29**(2): p. 119-67.
302. Ponting, C.P. and R.R. Russell, The natural history of protein domains. *Annu Rev Biophys Biomol Struct*, 2002. **31**(1): p. 45-71.
303. Eschmeyer, W.N., in *Annotated Checklists of Fishes*. 2003, Calif. Acad . Sci.
304. Tree of life web project, D. Maddison, Editor. 2004.
305. Robinson-Rechavi, M., et al., Euteleost fish genomes are characterized by expansion of gene families. *Genome Res*, 2001. **11**(5): p. 781-8.
306. Waterston, R.H., et al., Initial sequencing and comparative analysis of the mouse genome. *Nature*, 2002. **420**(6915): p. 520-562.
307. Okazaki, Y., et al., Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, 2002. **420**(6915): p. 563-573.
308. Henikoff, S., et al., Gene families: The taxonomy of protein paralogs and chimeras. *Science*, 1997. **278**(5338): p. 609-14.
309. Lynch, M. and J.S. Conery, The evolutionary fate and consequences of duplicate genes. *Science*, 2000. **290**(5494): p. 1151-5.
310. Gilligan, P., S. Brenner, and B. Venkatesh, *Fugu* and human sequence comparison identifies novel human genes and conserved non-coding sequences. *Gene*, 2002. **294**(1-2): p. 35-44.
311. Thomas, J.W., et al., Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 2003. **424**(6950): p. 788-93.
312. O'Brien, S.J., E. Eizirik, and W.J. Murphy, Genomics: On choosing mammalian genomes for sequencing. *Science*, 2001. **292**(5525): p. 2264-2266.
313. Hedges, S.B. and L.L. Poling, A molecular phylogeny of reptiles. *Science*, 1999. **283**(5404): p. 998-1001.
314. Boardman, P.E., et al., A comprehensive collection of chicken cDNAs. *Curr Biol*, 2002. **12**(22): p. 1965-9.
315. Copley, R.R., I. Letunic, and P. Bork, Genome and protein evolution in eukaryotes. *Curr. Opin. Chem. Biol.*, 2002. **6**(1): p. 39-45.
316. Hedges, S.B., The origin and evolution of model organisms. *Nat Rev Genet*, 2002. **3**(11): p. 838-49.
317. Carroll, R., Early land vertebrates. *Nature*, 2002. **418**(6893): p. 35-6.
318. Bly, J.E. and L.W. Clem, Temperature-mediated processes in teleost immunity: In vitro immunosuppression induced by in vivo low temperature in channel catfish. *Vet Immunol Immunopathol*, 1991. **28**(3-4): p. 365-77.
319. Dehal, P., et al., The draft genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. *Science*, 2002. **298**(5601): p. 2157-2167.
320. Cameron, R.A., et al., A sea urchin genome project: Sequence scan, virtual map, and additional resources. *Proc Natl Acad Sci U S A*, 2000. **97**(17): p. 9514-8.

321. Cameron, C.B., J.R. Garey, and B.J. Swalla, Evolution of the chordate body plan: New insights from phylogenetic analyses of deuterostome phyla. *Proc Natl Acad Sci U S A*, 2000. **97**(9): p. 4469-74.
322. Satou, Y., et al., A cDNA resource from the basal chordate *Ciona intestinalis*. *Genesis*, 2002. **33**(4): p. 153-4.
323. Shida, K., et al., Hemocytes of *Ciona intestinalis* express multiple genes involved in innate immune host defense. *Biochemical and Biophysical Research Communications*, 2003. **302**(2): p. 207-218.
324. Wada, H. and N. Satoh, Details of the evolutionary history from invertebrates to vertebrates, as deduced from the sequences of 18s rDNA. *Proc Natl Acad Sci U S A*, 1994. **91**(5): p. 1801-4.
325. Holland, L.Z. and J.J. Gibson-Brown, The *Ciona intestinalis* genome: When the constraints are off. *Bioessays*, 2003. **25**(6): p. 529-32.
326. Rubin, G.M., et al., Comparative genomics of the eukaryotes. *Science*, 2000. **287**(5461): p. 2204-15.
327. Loewen, M.C., et al., The ice-binding site of sea raven antifreeze protein is distinct from the carbohydrate-binding site of the homologous C-type lectin. *Biochemistry*, 1998. **37**(51): p. 17745-53.
328. Koonin, E.V., L. Aravind, and A.S. Kondrashov, The impact of comparative genomics on our understanding of evolution. *Cell*, 2000. **101**(6): p. 573-6.
329. Loots, G.G., et al., Rvista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res*, 2002. **12**(5): p. 832-839.
330. Hinegardner, R. and D.E. Rosen, Cellular DNA content and the evolution of teleostean fishes. *Am Nat*, 1972. **106**(951): p. 621-644.
331. Heierhorst, J., K. Lederis, and D. Richter, Presence of a member of the Tc1-like transposon family from nematodes and *Drosophila* within the vasotocin gene of a primitive vertebrate, the pacific hagfish *Eptatretus stouti*. *Proc Natl Acad Sci U S A*, 1992. **89**(15): p. 6798-802.
332. Hartl, D.L., A.R. Lohe, and E.R. Lozovskaya, Modern thoughts on an ancient mariner: Function, evolution, regulation. *Annu Rev Genet*, 1997. **31**: p. 337-58.
333. Vinogradov, A.E., Intron-genome size relationship on a large evolutionary scale. *J Mol Evol*, 1999. **49**(3): p. 376-84.
334. Rokas, A. and P.W.H. Holland, Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol*, 2000. **15**(11): p. 454-459.
335. Venkatesh, B., Y. Ning, and S. Brenner, Late changes in spliceosomal introns define clades in vertebrate evolution. *Proc Natl Acad Sci U S A*, 1999. **96**(18): p. 10267-71.
336. Cammarata, M., et al., A serum fucoselectin isolated and characterized from sea bass *Dicentrarchus labrax*. *Biochim Biophys Acta*, 2001. **1528**(2-3): p. 196-202.
337. Chevassus, B., Hybridization in fish. *Aquaculture*, 1983. **33**(1-4): p. 245-262.
338. Smale, S.T. and J.T. Kadonaga, The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, 2003: p. 449-479.
339. Patthy, L., Genome evolution and the evolution of exon-shuffling-a review. *Gene*, 1999. **238**(1): p. 103-14.
340. Hehl, R. and E. Wingender, Database-assisted promoter analysis. *Trends Plant Sci*, 2001. **6**(6): p. 251-5.

341. Drickamer, K., Evolution of Ca(2+)-dependent animal lectins. *Prog Nucleic Acid Res Mol Biol*, 1993. **45**: p. 207-32.
342. Brenner, S., et al., Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature*, 1993. **366**(6452): p. 265-8.
343. Aparicio, S. and S. Brenner, How good a model is the *Fugu* genome? [letter; comment]. *Nature*, 1997. **387**(6629): p. 140.
344. Venkatesh, B., P. Gilligan, and S. Brenner, *Fugu*. A compact vertebrate reference genome. *FEBS Lett*, 2000. **476**(1-2): p. 3-7.
345. Aparicio, S., et al., Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc Natl Acad Sci U S A*, 1995. **92**(5): p. 1684-1688.
346. Brunner, B., et al., Genomic structure and comparative analysis of nine *Fugu* genes: Conservation of synteny with human chromosome xp22.2-p22.1. *Genome Res*, 1999. **9**(5): p. 437-48.
347. Elgar, G., et al., Generation and analysis of 25 Mb of genomic DNA from the pufferfish *Fugu rubripes* by sequence scanning. *Genome Res*, 1999. **9**(10): p. 960-71.
348. Trower, M.K., et al., Conservation of synteny between the genome of the pufferfish (*Fugu rubripes*) and the region on human chromosome 14 (14q24.3) associated with familial alzheimer disease (ad3 locus) *Proc Natl Acad Sci U S A*, 1996. **93**(4): p. 1366-9.
349. McLysaght, A., et al., Estimation of synteny conservation and genome compaction between pufferfish (*Fugu*) and human. *Yeast*, 2000. **17**(1): p. 22-36.
350. Grutzner, F., et al., Four-hundred million years of conserved synteny of human Xp and Xq genes on three tetraodon chromosomes. *Genome Res*, 2002. **12**(9): p. 1316-1322.
351. Crnogorac-Jurcevic, T., et al., *Tetraodon fluviatilis*, a new puffer fish model for genome studies. *Genomics*, 1997. **41**(2): p. 177-84.
352. Crollius, H.R., et al., Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. *Genome Res*, 2000. **10**(7): p. 939-49.
353. Crollius, H.R., et al., Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence *Nat Genet*, 2000. **25**(2): p. 235-8.
354. Kristiansen, T.Z. and A. Pandey, A database of transcriptional start sites. *Trends Biochem Sci*, 2002. **27**(4): p. 174.
355. Lemon, B. and R. Tjian, Orchestrated response: A symphony of transcription factors for gene control. *Genes Dev*, 2000. **14**(20): p. 2551-2569.
356. Locker, J., et al., Definition and prediction of the full range of transcription factor binding sites--the hepatocyte nuclear factor 1 dimeric site. *Nucl Acids Res*, 2002. **30**(17): p. 3809-3817.
357. Nobrega, M.A., et al., Scanning human gene deserts for long-range enhancers. *Science*, 2003. **302**(5644): p. 413-.
358. Ureta-Vidal, A., L. Ettwiller, and E. Birney, Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet*, 2003. **4**(4): p. 251-62.
359. Muller, F., P. Blader, and U. Strahle, Search for enhancers: Teleost models in comparative genomic and transgenic analysis of cis regulatory elements. *Bioessays*, 2002. **24**(6): p. 564-72.
360. Clark, M.S., S.F. Smith, and G. Elgar, Use of the japanese pufferfish (*Fugu rubripes*) in comparative genomics. *Marine Biotechnology*, 2001. **3**: p. S130-S140.

361. Kondrashov, A.S., Comparative genomics and evolutionary biology. *Curr Opin Genet Dev*, 1999. **9**(6): p. 624-9.
362. Bonneau, R. and D. Baker, Ab initio protein structure prediction: Progress and prospects. *Annu Rev Biophys Biomol Struct*, 2001. **30**(1): p. 173-189.
363. Burley, S.K., An overview of structural genomics. *Nat Struct Biol*, 2000. **7 Suppl**: p. 932-4.
364. Elgavish, S. and B. Shaanan, Lectin-carbohydrate interactions: Different folds, common recognition principles. *Trends Biochem Sci*, 1997. **22**(12): p. 462-7.
365. Drickamer, K., Increasing diversity of animal lectin structures. *Curr Opin Struct Biol*, 1995. **5**(5): p. 612-6.
366. Rini, J.M. and Y.D. Lobsanov, New animal lectin structures. *Curr Opin Struct Biol*, 1999. **9**(5): p. 578-584.
367. Marti-Renom, M.A., et al., Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 2000. **29**(1): p. 291-325.
368. Chothia, C. and A.M. Lesk, The relation between the divergence of sequence and structure in proteins. *Embo J*, 1986. **5**(4): p. 823-6.
369. Vitkup, D., et al., Completeness in structural genomics. *Nat Struct Biol*, 2001. **8**(6): p. 559-66.
370. Bianchet, M.A., et al., A novel fucose recognition fold involved in innate immunity. *Nat Struct Biol*, 2002. **9**(8): p. 628-34.
371. Potterton, E., et al., The CCP4 molecular-graphics project. *Acta Crystallogr D Biol Crystallogr*, 2002. **58**(Pt 11): p. 1955-7.
372. Rini, J.M., X-ray crystal-structures of animal lectins. *Curr Opin Struct Biol*, 1995. **5**(5): p. 617-621.
373. Wright, C.S., New folds of plant lectins. *Curr Opin Struct Biol*, 1997. **7**(5): p. 631-6.
374. Bouckaert, J., et al., Novel structures of plant lectins and their complexes with carbohydrates. *Curr Opin Struct Biol*, 1999. **9**(5): p. 572-577.
375. Basu, A. and D. Kahne, Overcoming degeneracy in carbohydrate recognition. *Angew Chem Int Ed Engl*, 2003. **42**(22): p. 2504-2506.
376. Springer, G.F. and P.R. Desai, The immunochemical requirements for specific activity and the physicochemical properties of eel anti-human blood-group H(O) 7 S globulin. *Vox Sang*, 1970. **18**(6): p. 551-4.
377. Staudacher, E., et al., Fucose in N-glycans: From plant to man. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1999. **1473**(1): p. 216-236.
378. Stevenson, G., et al., Organization of the *Escherichia coli* K-12 gene cluster responsible for production of the extracellular polysaccharide colanic acid. *J Bacteriol*, 1996. **178**(16): p. 4885-4893.
379. Baldus, S.E., et al., Characterization of the binding specificity of *Anguilla anguilla* (AAA) in comparison to *Ulex europaeus* agglutinin I (UEA-I). *Glycoconjugate Journal*, 1996. **13**: p. 585-590.
380. Silva, J.J.R.F.d. and R.J.P. Williams, *The biological chemistry of the elements : The inorganic chemistry of life*. 2nd ed. 2001, Oxford ; New York: Oxford University Press. xvii, 575.
381. McPhalen, C.A., N.C. Strynadka, and M.N. James, Calcium-binding sites in proteins: A structural perspective. *Adv Protein Chem*, 1991. **42**: p. 77-144.
382. Bezkorovainy, A., G.F. Springer, and P.R. Desai, Physicochemical properties of the eel anti-human blood-group H(O) antibody. *Biochemistry*, 1971. **10**(20): p. 3761-4.

383. Horejsi, V. and J. Kocourek, Studies on lectins: XXXVI. Properties of some lectins prepared by affinity chromatography on o-glycosylation polyacrylamide gels. *Biochimica et Biophysica Acta*, 1978. **538**(2): p. 299-315.
384. Weis, W.I. and K. Drickamer, Trimeric structure of a C-type mannose-binding protein. *Structure*, 1994. **2**(12): p. 1227-40.
385. Lee, R.T., et al., Multivalent ligand binding by serum mannose-binding protein. *Arch Biochem Biophys*, 1992. **299**(1): p. 129-36.
386. Holm, L. and C. Sander, Dali: A network tool for protein structure comparison. *Trends Biochem Sci*, 1995. **20**(11): p. 478-480.
387. Macedo-Ribeiro, S., et al., Crystal structures of the membrane-binding c2 domain of human coagulation factor v. *Nature*, 1999. **402**(6760): p. 434-9.
388. Firbank, S.J., et al., From the cover: Crystal structure of the precursor of galactose oxidase: An unusual self-processing enzyme. *Proc Natl Acad Sci U S A*, 2001. **98**(23): p. 12932-12937.
389. Wendt, K.S., et al., Crystal structure of the apc10/doc1 subunit of the human anaphase- promoting complex. *Nat Struct Biol*, 2001. **8**(9): p. 784-8.
390. Marintchev, A., et al., Solution structure of the single-strand break repair protein XRCC1 N-terminal domain. *Nat Struct Biol*, 1999. **6**(9): p. 884-93.
391. Coulson, A.F. and J. Moult, A unfold, mesofold, and superfold model of protein fold use. *Proteins*, 2002. **46**(1): p. 61-71.
392. Murzin, A.G., et al., Scop: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 1995. **247**(4): p. 536-40.
393. Kinch, L.N. and N.V. Grishin, Evolution of protein structures and functions. *Curr Opin Struct Biol*, 2002. **12**(3): p. 400-408.
394. Bork, P. and R.F. Doolittle, *Drosophila* kelch motif is derived from a common enzyme fold. *J Mol Biol*, 1994. **236**(5): p. 1277-82.
395. Boraston, A.B., et al., Structure and ligand binding of carbohydrate-binding module CSCBM6-3 reveals similarities with fucose-specific lectins and "galactose-binding" domains. *J Mol Biol*, 2003. **327**(3): p. 659-669.
396. Baumgartner, S., et al., The discoidin domain family revisited: New members from prokaryotes and a homology-based fold prediction. *Protein Sci*, 1998. **7**(7): p. 1626-31.
397. Rosen, S.D., et al., Developmentally regulated, carbohydrate-binding protein in *Dictyostelium discoideum*. *Proc Natl Acad Sci U S A*, 1973. **70**(9): p. 2554-7.
398. Poole, S., et al., Sequence and expression of the discoidin i gene family in *Dictyostelium discoideum*. *J Mol Biol*, 1981. **153**(2): p. 273-89.
399. Zwaal, R.F., P. Comfurius, and E.M. Bevers, Lipid-protein interactions in blood coagulation. *Biochim Biophys Acta*, 1998. **1376**(3): p. 433-53.
400. Lee, C.C., et al., Crystal structure of the human neuropilin-1 b1 domain. *Structure*, 2003. **11**(1): p. 99-108.
401. Vogel, W., Discoidin domain receptors: Structural relations and functional implications. *FASEB J*, 1999. **13**(9001): p. 77-82.
402. Vogel, W., et al., The discoidin domain receptor tyrosine kinases are activated by collagen. *Mol Cell*, 1997. **1**(1): p. 13-23.
403. Taubes, G., Cardiovascular disease. Does inflammation cut to the heart of the matter? *Science*, 2002. **296**(5566): p. 242-5.

- 404. Secombes, C.J., et al., Cytokines and innate immunity of fish. *Dev Comp Immunol*, 2001. **25**(8-9): p. 713-23.
- 405. Matsushita, M., et al., A novel human serum lectin with collagen- and fibrinogen-like domains that functions as an opsonin. *J Biol Chem*, 1996. **271**(5): p. 2448-54.
- 406. Bharadwaj, D., et al., Serum amyloid p component binds to fc gamma receptors and opsonizes particles for phagocytosis. *J Immunol*, 2001. **166**(11): p. 6735-6741.
- 407. Kuhlman, M., K. Joiner, and R.A. Ezekowitz, The human mannose-binding protein functions as an opsonin. *J Exp Med*, 1989. **169**(5): p. 1733-45.
- 408. Ghiran, I., et al., Complement receptor 1/CD35 is a receptor for mannan-binding lectin. *J Exp Med*, 2000. **192**(12): p. 1797-1808.
- 409. McCaffrey, A.P., et al., RNA interference in adult mice. *Nature*, 2002. **418**(6893): p. 38-9.